*

Great exploration. Graphics a little weak but mainly excus fault not yours. You discover some really interesting relationships

Stat 480, Project #1

February 20, 2007

Introduction to the Data

The primary interest in this data set pertains to earnings and the relationship various factors have on AdSense earnings. The data set provided by AdMoolah.com provides, in addition to earnings, six pieces of additional data: date, website category, primary language of the website, the website's PageRank, number of visitors to the website, and page views. Each piece of data has the possibility of explicating the cause—or at least the correlation—of higher Google AdSense earnings. Instead of merely plotting earnings against other variables, we can first rationalize the relationships we might expect. These rationalizations will help us ask more poignant questions for our statistical analysis. In the following text we will pose some questions and describe the process of collecting data.

? poignant is NOT the word you are looking for

Questions:

AdSense earnings are loosely based on the number of clicks on an advertisement. We would expect more visitors will lead to a greater number of ad clicks, thereby increasing AdSense earnings. Therefore, our first question is, *do more visitors imply higher AdSense earnings?* However, it is unclear if page views are positively or negatively related to earnings. More page views imply the same set of people consistently returning to the same web page. Those persons may have either learned to ignore the ads entirely or may find the advertisements that match their interests. Thus, our second question is, *how are page views related to earnings?* ✓

PageRank is a system that ranks websites based on popularity and reputability. These relationships are then assigned a score—from zero to ten—based on links from other websites. Page rank is essentially a system of reputability points; a higher score means the websites is more reputable. Therefore, it seems very plausible that visitors will trust advertisements on a reputable website more than a less reputable website. Our third question is, *does a higher PageRank imply higher AdSense earnings?*

The website's primary language might also impact earnings. We can therefore ask the hypothetical question: *does having an English website increase AdSense earnings?* Similarly, the website category might also influence earnings. Google may have a harder time matching a personal blog—which might vary in topic day-to-day—with relevant ads. A more focused website, such as one exclusively about travel, might serve better ads that interest the reader. That leads us to our last question: *how does the website category affect earnings?*

These questions can help us play an advisor role to webmasters. We might be able to advise webmasters to work on establishing a higher page rank to increase AdSense earnings. Likewise, we might advise a webmaster to switch from Spanish to an English-

C&Q 9 ~~10~~/10 ← because you didn fix poigna
G&T 8/10
Find. 18/20
Conc. 18/20

only blog. Each one of these questions is geared to establishing the causal relationships of higher earnings, but they are also geared to being a consultant to web site owners.

Data Collection

The database of AdMoolah's AdSense earnings data is available for the public to query. There is no complete listing of the data; rather one must search for data that meets specific criteria. Many of the data are incomplete, however, sometimes omitting PageRank or data on number of visitors or views. None of the data is missing a date, however, since this value is added when the user inputs her AdSense data. It is possible, then, to select the complete range of dates and gain access to all of the available AdSense data. However, that is not the tactic that was employed.

The method used was to select the data with a valid PageRank listed (0-10), a valid number of monthly visitors (0 to 999999999), and a valid number of monthly views (0 to 999999999). The questions analyzed in this paper will require all of the said data, and there are two reasons why we would like the data to be complete. First, we would like to have the same sample available for analysis with each question. It is possible that by chance, data with omitted values could exhibit different trends than complete data. This would lead to inconsistent results over our set of questions. Secondly, omitted values may be indicative of a user who is not very familiar with his website. We would prefer to analyze data that was inputted by webmasters very knowledgeable of their site. If a user does not know how many visitors or page views his site receives, the rest of the inputted data is likely not as reliable.

✓ good explanation

The mechanics of the data collection were straight-forward. After querying for data, it was displayed in an HTML table which was copied and pasted into Excel. Coercing the data into a usable form was slightly harder, however. First, all formatting had to be removed; this was done by "pasting values" into a new sheet. Secondly, earnings had to be turned into a single number. Some data points indicated that the user had several websites; this piece of information was ignored, however, since the data indicated that the sites shared the same category, language, and PageRank. Finally, the category names were shortened and simplified. By eliminating the spaces and punctuation, they will be easier to reference during analysis. The category names were also disaggregated so that the data could be analyzed by the broad or narrow categories (e.g. Computers and Computers: Internet).

✓ good

Question 1, *do more visitors imply higher AdSense earnings?*

AdSense earnings are loosely based on the number of clicks on an advertisement. We would expect more visitors will lead to a greater number of ad clicks, thereby increasing AdSense earnings. In order to investigate this hypothesis the data was analyzed and grouped in a variety of ways using Pivot Plots and graphs. The number of visitors was grouped by 100, 1,000, 10,000, and then 100,000.

When the visitors were sectioned off by 10,000's almost all of the count of websites was in the first section (0-10,000 visitors). By sectioning it off in this manner it greatly skewed the results of the analysis. The results claimed that that first group (of 0-10,000 visitors) had the highest AdSense earnings. The reason for this inaccurate analysis was simply because most of the sites were falling into this first subset (0-10,000 visitors), and the remaining subsets (10,000-20,000 visitors, 20,000-30,000 visitors…) were only based on one or two sites. The number of websites needed to be more evenly distributed between subsets in order to get a more reliable result.

The most accurate representation of the data seems to be when the number of visitors is sectioned off by 100's (0-100 visitors, 100-200 visitors…). In this sequence each subsection had approximately the same number of websites. The plot of this data showed a gradual curved increase in the amount of AdSense earnings as the number of visitors increased. This means that the more visitors a website was able to bring in, the greater the number in ad clicks, and therefore a greater amount of AdSense earnings.

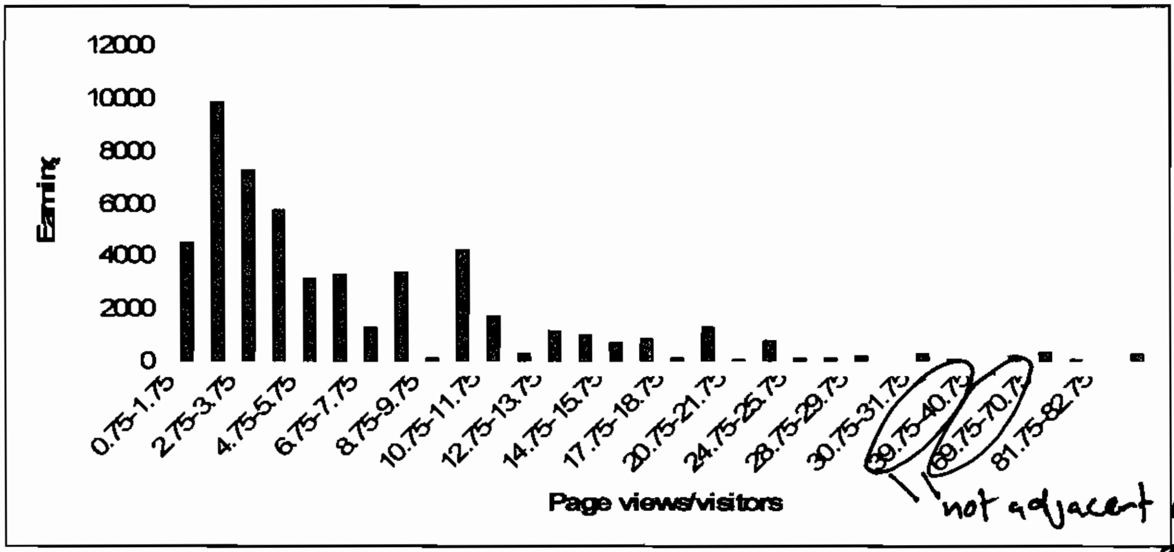
do you have a figure showing thi

It seems that in this case our initial hypothesis was confirmed. "We would expect more visitors will lead to a greater number of ad clicks, thereby increasing AdSense earnings." This conclusion makes logical sense, the more people in the site the more likely one of them is to click on one of the advertisements. The click then brings greater AdSense earnings. Next, we need to examine if this same concept holds through for page views for each website.

Question 2, *how are page views related to earnings?*

More page views imply the same set of people consistently returning to the same web page. We need to determine whether page views are positively or negatively related to earnings. Have these people learned to ignore the ads entirely or are the finding the advertisements that match their interests?

In order to solve this question the number of page views had to be divided by the number of visitors for each site. This will give the approximate page views per visitor. The higher this number is the more times the visitors went back to the site.



not adjacent ranges! excel is tricking you + hard to read

As shown in the above graph there seems to be a negative relationship between the number of times the page is viewed and the amount of earnings. Therefore, just because a visitor returned to the site more then once, does not mean the advertisement was more effective the second or third time the person came in.

One possible conclusion drawn from this analysis is that visitors had already seen the ads once, and if they didn't click on them the first time what incentive did they have to do so the second time? They may have simply become used to the ads and no longer paid attention to them.

Interesting speculation

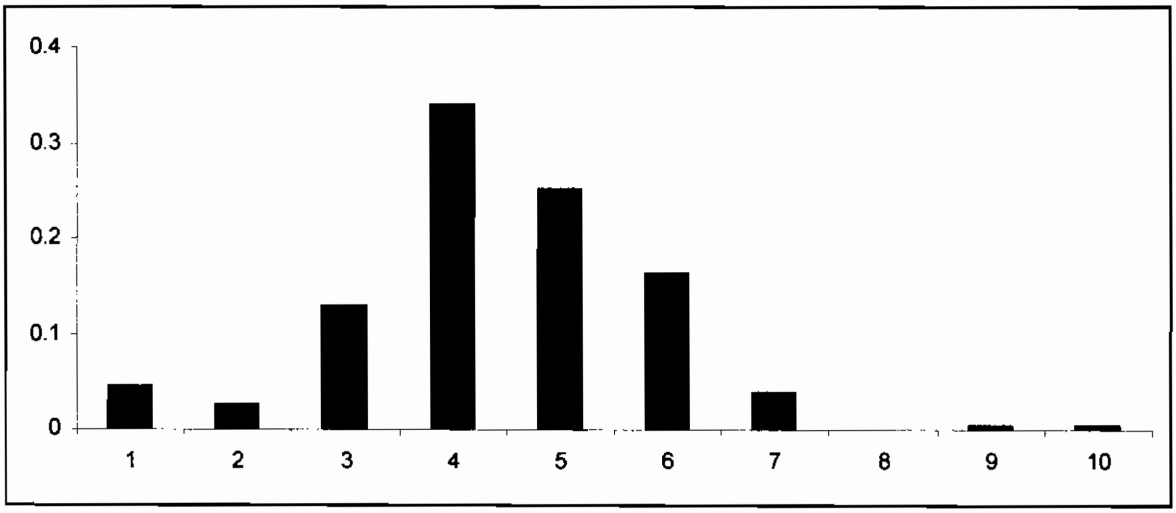
Therefore companies posting their ads through AdMoolah should be more concerned with posting their ads on a site that simply draws in large numbers of visitors relating to their product or service, but not necessarily brings them back again.

Question 3, *does a higher page rank imply higher AdSense earnings?*

PageRank is the method Google uses to order search results. It is an algorithm developed by Larry Page (hence, *Page*Rank) and Sergey Brin while at Stanford. In the algorithm Google describes as "democratic," PageRank gives each website a score based on the number and quality of links pointing to that site. A higher score indicates that a website is frequently linked to, or linked to by prominent websites. In short, PageRank intends to be a measure of popularity and reputability of a website.

Because of this, we would expect companies to pay a premium to have their advertisements appear on websites with high PageRanks. There will be more visitors to these sites, and the visitors would likely have more trust in the reputability of the advertising business partners than with other sites. Clearly, then, we would expect Google's payments for advertisement hosting to be proportional to the PageRank of the site in question. ✓

To begin this inquiry, first consider the distribution of PageRanks represented in our data:



Out of the sample of 239 sites, there is only one site with a PageRank of 10, and only one site with a score of 9. The distribution peaks at 4 and looks roughly symmetrical, although there is more concentration in the lower tail than the higher tail.

It is unlikely that this distribution is representative of Google's AdSense client base. Google is interested in giving its advertisements high visibility, so in reality a good portion of its clients will have high PageRanks (say, between 7 and 10). Conversely, Google likely has relatively few AdSense clients with a small PageRank (between 1 and 3); these smaller sites will have few page views and even fewer visitors, so advertisements would not be as effective. The number of mid-ranked (between 4 and 6) sites is likely to have the highest frequency in the program, as these sites are much more common on the Internet (information on the PageRank distribution of all websites was not available, but it is likely similar to an exponential distribution).

I don't thin this is true. Goo doesn't de who uses adsense on their web

The discrepancy between our sample distribution of PageRanks and the true distribution is likely caused by self-selection in data reporting. The data at AdMoolah is all voluntarily reported by webmasters. Because of this, small sites will be over-reported compared to larger sites. Small websites are typically a hobby of the webmaster, and the object of the master's pride; thus the site master would be happy to discuss and share data about her website. With large sites, however, it takes many fulltime employees to manage them. These employees are likely not as apt to discuss their site on small websites like AdMoolah, and additionally, they may be prevented from doing so through company policy.

Good point

The following table shows the effect of PageRank on various levels of earnings. The first column measures average total monthly AdSense earnings. The second measures average monthly earnings per page view, and the third measures average monthly earnings per visitor. The last column measures the average earnings per view per visitor – that is, the amount of money earned divided by the average number of pages each visitor loads. The PageRank values are aggregated into the three categories discussed above. Without aggregation, there are very few sites per PageRank value and there is much more noise; because of this, only the aggregate table is shown below.

PageRank	$/mo	($/mo)/view	($/mo)/visitor	($/mo)/(view/visitor)
Low (1-3)	237.01	0.0091	0.0310	64.03
Medium (4-6)	259.80	0.0121	0.0272	62.75
High (7-10)	1507.83	0.0003	0.0029	59.50
Total	312.66	0.0109	0.0268	62.85

An interesting pattern emerges from these tables: even though a higher PageRank is associated with higher monthly earnings, a site with a high PageRank earns substantially less money per click and per visitor than a site with a low PageRank. One possible explanation for this is that high-PageRank sites have more-frequent but shorter visits than low-PageRank sites, since most prominent websites are used for a quick check of news, search listings, or other data. If this were the case, the visitors who spend little time per page view would also spend little time looking at advertisements, so the view is not as valuable to the advertiser.

great questions

Across all values of PageRank, however, the monthly earnings per visitor-view is roughly constant. A likely explanation for this is that the views per visitor are a good indicator of how much time a user looks at a particular website. Instead of paying based off how many times a computer loads a web page into memory, an advertiser would be much more interested in paying based off how long a user looks at a particular website containing advertisements.

Given this, we can say that the value of monthly views and visitors to advertisers does not increase with PageRank; in fact, the opposite seems to hold. The time spent at the website, proxied by the average number of views per visitor, seems to be constant over the PageRank categories. Thus the earnings differential between PageRank groups must be explained by differences of average views per visitors (and indeed, the data support this: the average number of views per visitor is 8.98 for low PageRank, 9.80 for medium, and 15.75 for high).

✓

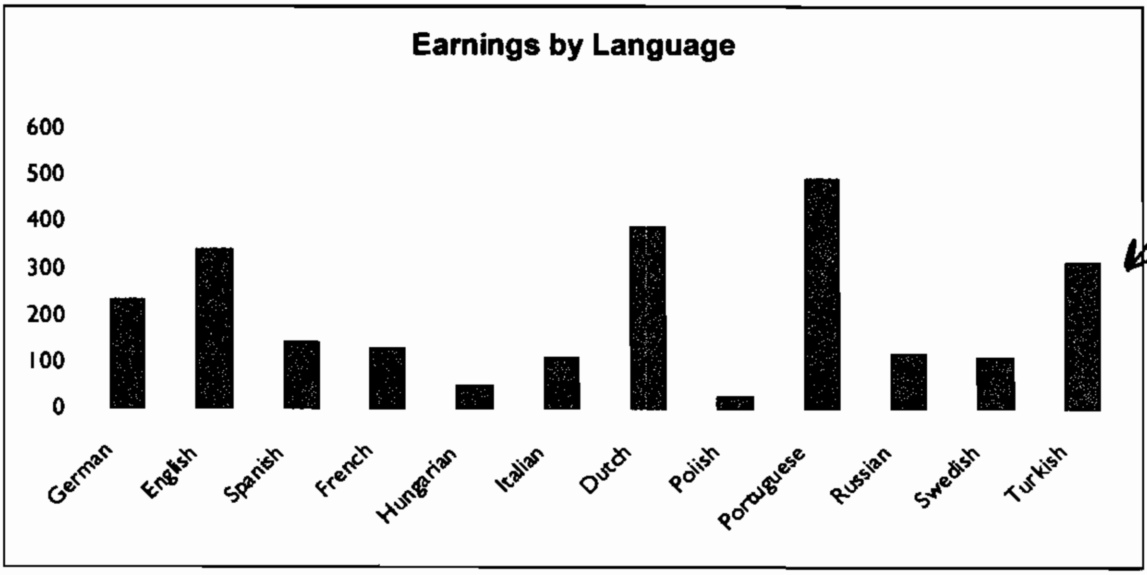
Question 4, : does having an English website increase AdSense earnings?

Although English is not the largest primary language, approximations[1] of the most popular websites show English-based websites occupy 54 of the top 100 most popular websites. A potential explanation for this is English's status as the largest secondary language and its use to communicate across cultures. Because English-based websites draw a higher proportion of web traffic, webmasters might be behooved to author their respective web pages in English. The goal of this section is to qualitatively test this notion.

← Great!

The AdMoolah data set contains 174 observations reported as English websites—the largest group in the sample. The second largest group is German-based websites with 37 observations, subsequent groups are the Dutch (8 observatons), Italian, and Turkish both with 4 observations. Compared with the global distribution of languages, English and German is over represented while Mandarin, which has the largest group of speakers, is not represented at all in our data set. The absence of Mandarin is especially conspicuous since the Chinese language is the second most popular language in the top 100 websites on the internet.

Let us first consider the average earnings by country shown in the figure below. English receives an average of $343 in AdSense revenue. Only four other languages have average revenue over $200: German, Dutch, Portuguese, and Turkish. Both Dutch and Portuguese have more average earnings than English websites with $388 and $494, respectively.



← how is this chart ordered?

The average Portuguese earnings rely upon a single observation. The other Portuguese samples had incomplete information and were dropped from our analysis. The sole Portuguese observation is a moderately successful website with a PageRank of 5, 72,779 visitors, and 488,320 page views. The average earnings for sites with a PageRank of 5 is $304. The Portuguese website earned $494, $190 more than the average, but well within the $406 standard deviation for that group of websites. While it is possible the

[1] Approximations were calculated from Alexa.com's Global Top 500, which was accessed on February 12, 2007.

additional $190 can be attributed to the Portuguese language, it is not feasible to distinguish between a language effect and statistical noise. Therefore, we cannot conclude there is an earnings benefit for websites written in Portuguese. ✓

Dutch websites average $494 in AdSense revenues, $151 more than English websites. Unlike the Portuguese, which was based on a single observation, there are 8 observations of Dutch websites. The table below compares several categories between English and Dutch websites.

	Average Number of Views	Average Page Rank	Earnings Per Visitor-View	Views Per Visitor
English	818, 830	4.4	67.5	8.3
Dutch	212,043	5.5	117.9	4.3

The Dutch websites appear more effective in terms of drawing more revenue per view for each visitor. This is despite the fact that English websites have a higher average number of views and even views per visitor. Dutch websites also have a higher average PageRank, which may explain the higher earnings. In either case, it does not appear that websites written in English receive an earnings premium. ✓

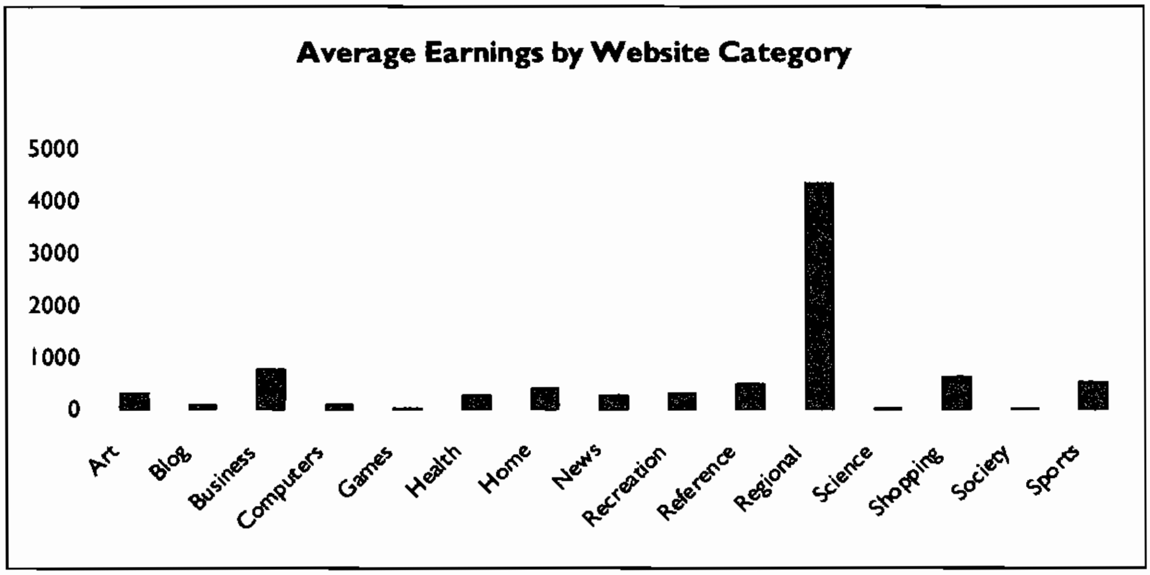
Question 5, *how does the website category affect earnings?*

Google's AdSense permits almost anyone to place ads on their website. As a result, a variety of websites take part in the program, such as personal websites hoping to recoup operating expenditures to larger revenue generating websites. These websites also cover a number of topics. In this specific data set there are 40 website categories.

For this section, Google's AdWords technology will impact our results. AdWords is used in conjunction with AdSense to determine which ads are placed on which websites. To effectively pair ads with websites, Google scans the website for specific words, such as computers, to estimate the website's content. Ads that fit these keywords are then placed through the AdSense service. Essentially, the AdWords technology helps increase ad clicks on webpages, thus increasing revenues for Google and its AdSense partners.

Since blogs tend to discuss a variety to topics, one would suspect the Google AdWords technology will be less effective in placing good advertisements. Therefore, one would also suspect blogs will have lower revenues. Other website categories, which tend to be more focused on a single topic, will raise more revenues.

good questions



Average earnings is highest for the regional website category ($4,362), followed by business ($777) and shopping ($630) categories. A closer look at the regional category reveals three observations; one is a regional website for Canada which lists gross earnings of $11,982, which happens to be the highest grossing website in the data set. The two other observations have revenues of $605 and $500, respectively.

There is additional corollary evidence that the regional Canadian website is a generally successful website. It receives around one million visitors and one-hundred million web views, suggesting it is updated often. Furthermore, it has a PageRank of 9 and earns $119 for each visitor-view—nearly twice the typical amount. In short, this website excels in every data category.

Excellent! Removed outli & recomputed values

If we were to remove the regional Canadian site, average earnings for the regional category would be $552 for the two remaining observations, placing third behind business and shopping. This suggests that regardless of the aberrant regional Canadian website, the regional category is associated with high earnings. Meanwhile, the blog category ($120) only earned more on average than the science ($47), games ($38), and society ($32).

Previously, we've found earnings per visitor-view have significant influence on gross earnings. Earnings per visitor-view are a proxy for the quality of readership. A higher ratio implies the visitor stays longer, reading the material instead of skimming. It is plausible certain categories attract higher readership quality, thus earning higher AdSense revenues.

probably wouldn't use such a value laden descript

To determine whether categories receive higher quality readers, we will first rank the categories by their average earnings. For instance, since regional has the highest average earnings, it receives a ranking of 1, followed by business, shopping, and so-forth. We then sort and rank each category by earnings per visitor-view. In this sample, sports has the highest earnings per visitor-view, therefore it receives a ranking of one.

Category	Rank of Average Earnings	Rank of Average Earnings per Visitor-View
Art	7	4

Blog	12	7
Business	2	6
Computers	11	13
Games	14	12
Health	9	9
Home	6	3
news	10	5
Recreation	8	11
Reference	5	8
Regional	1	10
Science	13	14
Shopping	3	2
Society	15	15
Sports	4	1
Correlation	63.93%	

Great novel analysis. A scatterplot would be useful here fo

The top three average earning website categories, regional, business, and shopping, rank tenth, sixth, and second, respectively. Overall, the correlation between average earnings and average earnings per visitor-view is positive, possibly implying certain website categories attract higher quality readers, thus earning higher AdSense revenues.

<u>Overall findings: *What really causes higher AdSense earnings?-Conclusion*</u>

As discussed earlier, the primary interest in this data set pertains to earnings and the relationship various factors have on AdSense earnings. The data set provided by AdMoolah.com provides, in addition to earnings, six pieces of additional data: date, website category, primary language of the website, the website's page rank, number of visitors to the website, and page views. Each piece of data has the possibility of explicating the cause—or at least the correlation—of higher Google AdSense earnings.

In the process of answering five questions, the reasons and correlations between the factors and of AdSense earning become apparent. The first logical conclusion that was introduced was that the more visitors that came to a site, the higher the AdSense revenues would be. This assumption, that more visitors would equal more clicks on an ad, held true. This made visitors the first essential item when figuring what is important for higher AdSense earnings.

The second question brought a very different conclusion. It seemed that there was a negative relationship between the number page views per visitor and the amount of earnings. Therefore, just because a visitor returned to the site more then once, does not mean the advertisement was more effective the second or third time the person came in. The second assumption did not help to better understand what causes higher AdSense earnings and it seems that getting the visitors to the site is important, not making them return. ✓

An interesting pattern seems to emerge in our third question. It seems that although a higher PageRank leads to higher monthly earnings, it also leads to substantially lower money per click and per visitor then a site with a lower PageRank. Given this, we can say that the value of monthly views and visitors to advertisers does not increase with PageRank; in fact, the opposite seems to hold. Therefore PageRank does not help to increase AdSense earnings either, in this case Google may want to choose to put their client's ads on websites with lower page ranks, it they want higher AdSense earnings for the ad.

The fourth assumption did not yield the expected results either. Although there is a far greater number of websites in English on the web, it would be expected that these sites would generate greater AdSense earnings. However the Dutch example proved this assumption wrong. The Dutch websites appear more effective in terms of drawing more revenue per view for each visitor. This is despite English websites having a higher average number of views and even views per visitor. Dutch websites also have a higher average PageRank, which may explain the higher earnings. In either case, it does not appear that websites written in English receive an earnings premium. In this case Google should choose to put their client's advertisements on Dutch websites over English ones. However with a larger proportion of people speaking English as a first or second language it may not help hit the company's target audience. ✓

The last question deals with what category the websites and ads are in, and which of these helps raise AdSense earnings. In the preliminary analysis of the data; business, shopping and regional websites seem to average highest AdSense earnings. However when a different approach was taken using quality readers over readers who just skim the site the results changed placing sports, shopping and home on the top. Overall, the correlation between average earnings and average earnings per visitor-view is positive, possibly implying certain website categories attract higher quality readers, thus earning higher AdSense revenues. However through both tests the category of shopping seemed to stay in the top three, indicating that whether quality readers are needed or not, shopping sites would be a superior place for Google to place an Ad. ✓

Overall, what should Google do to get higher AdSense earnings for a particular ad? They should pick a site (one with low PageRank) that gets lots of quality Dutch visitors who want to shop to come to their website. But of course the point of Google is to place Ad's on sites where they will be most effective, and different Ad's need to be on different sites to be truly effective with their respective audiences. Still the questions and analyses presented above will help get a better idea of what aspects of Google's AdSense earnings are caused by, and how they could potentially be changed.

However discrepancies may appear between these analyses and the truth, this is due to self-selection and data reporting into the AdMoolah website. The data at AdMoolah is all voluntarily reported by webmasters. Because of this, small sites will be over-reported compared to larger sites. Small websites are typically a hobby of the webmaster, and the object of the master's pride; thus the site master would be happy to discuss and share data about her website. With large sites, however, it takes many fulltime employees to manage them. These employees are likely not as apt to discuss their site on small websites like AdMoolah, and additionally, they may be prevented from doing so through company policy. ✓

Good point,
a discussion of a
better data collection
process would be nice.