ANLY 506 Quiz 2_

Due Jul 21 at 11pm **Points** 100 **Questions** 10

Available Jul 7 at 8:43pm - Jul 21 at 11:01pm 14 days **Time Limit** None

Instructions

Dataset used:

college Dataset.csv

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	36 minutes	100 out of 100

Score for this quiz: **100** out of 100

Submitted Jul 20 at 9:48pm

This attempt took 36 minutes.

Question 1

10 / 10 pts

Select the best answer below that names the principle or principles of analytic graphing. (These principles are important because they establish the foundation or reason for conducting analytic graphing.)

○ Content is king

○ To describe and document, and show multivariate data

Correct!

● All the above

○ To integrate evidence in order to make a convincing argument or story

○ To show comparisons and causality

Question 2

10 / 10 pts

The reason that we conduct exploratory graphing is to:

○ Understand data properties

○ Find patterns in data

○ Develop a personal understanding of the data

○ Find underlying structure in data

Correct!

● All the above

Question 3

10 / 10 pts

Things that compound the explanation of relationships, for example seasonal effects on average temperatures, are examples of Simpson’s Paradox and can be difficult to tease out of data.

Correct!

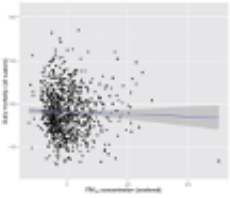
● True

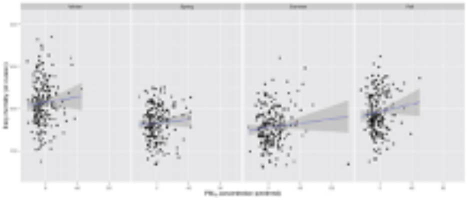
○ False

Question 4

10 / 10 pts

The figures below, the first of which appears to show that daily mortality decreases with increasing levels of pollution followed by seasonal plots of the same data which show the reverse i.e. daily mortality increases with increasing levels of pollution, provide an example of:





Correct!

● Simpson’s Paradox

○ None of the above

○ Chekhov’s Gun

○ Occam’s Razor

Question 5

10 / 10 pts

Let’s do some exploratory graphing. We’re going to use the College.csv dataset provided for you on Moodle. This dataset comes from Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani as part of their textbook ***An Introduction to Statistical Learning with Applications in R***. More information can be found online at: **https://cran.r-project.org/web/packages/ISLR/ISLR.pdf (https://cran.r-project.org/web/packages/ISLR/ISLR.pdf)**

The College dataset provides us with 777 observations of 19 variables. That is, it is a medium to large dataset. The variables include all types; numeric, integer, factor, etc. So, let’s first look at some summary or descriptive statistics. Select the best answer from the following regarding these summary statistics.

Hints: Don’t forget to put the dataset into your working directory for R. Then, read the dataset into R using the “read.csv” command. The dataset is on Moodle for you.

○ There are 777 private schools listed in the college dataset. The average number of applications (Apps) received by all schools annually is 1,558 and, on average, 1,110 are accepted (Accept); of those, an average of 434 actually enroll (Enroll). Annually there are an average of 9,990 out of state students (Outstate) across all schools listed.

○ None of the above.

Correct!

◉

There are 565 private schools listed in the college dataset. The average number of applications (Apps) received by all schools annually is 3,002 and, on average, 2,019 are accepted (Accept); of those, an average of 780 actually enroll (Enroll). Annually there are an average of 10,441 out of state students (Outstate) across all schools listed.

○

The average student to faculty ratio (S.F. Ratio) is 39.80. The average graduation rate (Grad. Rate) is 78 percent. And, the maximum number of students who completed high school in the top 10 percent of their graduating class (Top10perc) is 96.

Question 6

10 / 10 pts

Now that we've seen some of the summary statistics let's see if we can find make any comparisons. First let's look at the number of applicants to private versus public schools using boxplots.

Hint: You can use the R command

```
> boxplot(Apps ~ Private, college, ylim=c(0,25000), xlab="Applications to Private Schools (Yes/No)")
```

to do this. The purpose of the ylim is to cut off the extreme upper limit of the whiskers on the apps to private schools=No boxplot. Also, if applications are not to private schools they are to public schools.

○

Many more apps to public schools lie above the 75th percentile or upper quartile than there are for private schools.

○ Apps to public schools are on average twice as many as are made to private schools.

Correct!

● All the above.

○ The dispersion in the number of apps to public schools is roughly twice that of apps to private schools.

Question 7

10 / 10 pts

We can focus in on some more details. Let’s find out how many students who enrolled in private schools were in the top 10 percent of their graduating class.

Hints: To do this first subset (or divide out) the observations for students where Private = “Yes”. You can do this using the R command:

```
> private_apps = college[college$Private == "Yes", ]
```

Once you have the observations (or records) you want to examine then you can use the summary() command to look at the statistics.

○ There is not enough data available to determine this.

○ Of the students who enrolled in private schools roughly 96% were in the top 10 percent of their graduating high school class.

Correct!

● Of the students who enrolled in private schools just over 29% were in the top 10 percent of their graduating high school class.

○ None of the above.

Question 8

10 / 10 pts

Let’s look for some relationships in the data. First, let’s look for a relationship between the variables representing the number of applications (Apps) and the number of acceptances (Accept). Select the best answer from the following.

Hints: Just using the plot() command really won’t be sufficient to look at the data and find the best answer. You’ll want to differentiate between applications to private and public schools, perhaps by using different colors. To do this you’ll need to understand the order in which R assigns colors, i.e. R uses black first, then red, green, blue, cyan, magenta, yellow and last gray. And, you’ll want to limit the range for both variables to “zoom” into see the relationships. Also, to keep things simple we can use the syntax of “data$variable” so that for the college dataset to use the variable Apps you’d enter college$Apps. You could use the following R command to create a plot for use in exploring the relationship between number of applications and number of acceptances.

```
> plot(college$Apps, college$Accept, col=college$Private, ylim=c(0,
10000), xlim=c(0,15000))
```

This will assign different colors to Private=No first then to Private=Yes, and limit the x and y ranges. You can find the order of the levels for a variable by using the head command, e.g. head(college$Private).

○ There does not appear to be any relationship between the variables Apps and Accept.

○ None of the above.

○ There is not enough data to make this evaluation.

Correct!

● There appears to be a positive linear relationship regardless of whether public or private schools are involved. And, overall the number of acceptances at private schools appears to be lower.

Question 9

10 / 10 pts

Now do the same analysis for any relationship between the number of acceptances (Accept) and the number of actual enrollments (Enroll). Select the best answer from the following:

Hints: By analogy, the same as for the last question, Question #8. Don't forget remove the x and y range limits or adjust them as required.

○ There does not appear to be any relationship between the variables Accept and Enroll.

○ None of the above.

Correct!

●

There appears to be a positive linear relationship regardless of whether public or private schools are involved. And, overall the number of actual enrollments at private schools appears to be lower.

○ There is not enough data to make this evaluation.

Question 10

10 / 10 pts

1. There is one more technique we can use to explore our college data. We can look at the correlation between variables. Let's keep it simple and just consider the variables we've primarily been looking at from the college data, i.e. Apps, Accept, Enroll, Top10perc, and Outstate. Select the best answer from the following:

Hints: you can subset these variables into their own data frame, check to make sure the data frame correctly includes all variables; and, then run the cor() command one time for all of them as follows:

```
>subcollege<- data.frame(college$Apps, college$Accept,
college$Enroll, college$Top10perc,
college$Outstate)>str(subcollege)>cor(subcollege)
```

○ There appears to be inverse correlations between out-of-state students and enrollments as well as out-of-state students and acceptances.

Correct!

● All of the above

○

The strongest correlation appears to be between the number of college applications and the number of student acceptances. The second strongest correlation appears to be between the number of acceptances and enrollments.

○

The number of out-of-state students is least well correlated with the number of applications and acceptances.

Quiz Score: **100** out of 100