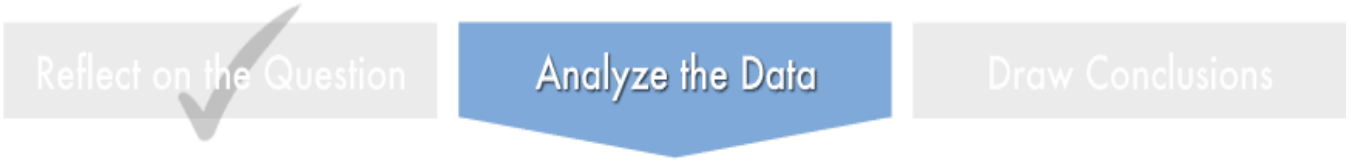
Analyze the Data



Primary Research Question

In 2012, which variable had the strongest linear relationship with Earnings: Ride Percentage or Cup Points?

Analysis

Let's break this analysis into the different steps that you will need to take to construct a complete answer. Be sure to:

1. Create a dataset which contains riders that participated in at least one event in 2012. Call the dataset **new_bull12**.
2. Make a histogram to visualize the distribution of Earnings for 2012.
3. Generate the appropriate descriptive statistics for this distribution.
4. Make a correlation matrix for Earnings12, RidePer12 and CupPoints12.
5. Plot a scatterplot for Earnings12 with each variable of interest. **Put Earnings12 on the y-axis.** Check for outliers.
6. Determine which variable has the strongest linear relationship with *Earnings12*.

problem

3/3 points (graded)

Earnings Distribution

1a. What is the **shape** of the Earnings distribution for 2012?

- (•) positively skewed ✔
- ( ) negatively skewed

1b. What was the **average** amount earned by a bull rider? (Choose the appropriate measure of center; report without a $ sign and round to the nearest whole number.)

147952 ✔

147952

1c. What was the **highest** amount earned by a bull rider? (Report without a $ sign and round to the nearest whole number.)

1464476 ✔

1464476

Submit You have used 1 of 1 attempt

problem

2/2 points (graded)

Make a Scatterplot of Earnings and Ride Percentage

2a. Does the scatterplot show a **linear** relationship?

Yes ✔

2b. What is the **correlation** of Earnings with Ride Percentage for 2012? (round to three decimal places)

.593 ✔

.593

Submit You have used 1 of 1 attempt

problem

2/2 points (graded)

Create a Scatterplot of Earnings and Cup Points

3a. Does the scatterplot show a **linear** relationship?

Yes ✔

3b. What is the **correlation** of Earnings with Cup Points for 2012? (report to three decimal places)

.657 ✔

.657

Submit You have used 1 of 1 attempt

---

problem

5/5 points (graded)

Outliers and Influential Points

An outlier can have a significant impact on the correlation coefficient. Sometimes it is important to remove these points to examine the size of this impact. Run this code to **identify** the extreme data value in Earnings:

```
# identify specific case
which(new_bull12$Earnings12 ==
max(new_bull12$Earnings12))
```

4a. The extreme earnings data point belonged to the rider that came in _____ Place in 2012. (Please spell your answer; do not use numerals.)

first ✔

4b. Where does this data point fall in the scatterplot? (**Make sure that Earnings12 is on the y-axis**)

- (•) Above the line ✔
- ( ) Below the line

On the line

Let's **remove** this data point from the dataset to assess what kind of impact, if any, it had on our correlation analysis. Run this code:

```
#Subset the data
nooutlier <- new_bull12[new_bull12$Earnings12 < 1000000
,]
```

Then **rerun** the correlation matrix and the scatterplots to see the difference. Make sure to use the new dataframe (nooutlier) that you just created.

4c. After removing the outlier, what was the **new correlation** of Earnings and Ride Percentage for 2012? (Round to three decimals)

.804 ✔

.804

4d. After removing the outlier, what was the **new correlation** of Earnings and Cup Points for 2012? (Round to three decimals)

.893 ✔

.893

4e. We would say that this data point was an **influential point** because it

- caused the underlying relationship to be non-linear.
- inflated the relationship between Earnings and the other variables.
- made the earnings of the other bull riders look less impressive than they really were.
- masked the strength of the relationships between Earnings and the other variables ✔

Submit You have used 1 of 1 attempt