Simple Analysis

Grant Esparza

September 6, 2018

Introduction

For this assignment I will be looking at records of political lobbyist activities in the city of Chicago. I obtained this data from the website Kaggle, which was orignally found on the City of Chicago's organization page. There were several datasets to choose from, however I decided to focus on the relationship between a lobbyist's compensation and how that affected the amount of money donated to political organizations. This information was found in two seperate files, `lobbyist_data-compensation.csv` and `lobbyist-data-contributions.csv`.

Data Preparation

```
## load libraries
library(ggplot2)
library(dplyr)
library(pander)

## Read data
comp_dat <- read.csv("lobbyist-data-compensation.csv")
contribute_dat <- read.csv("lobbyist-data-contributions.csv")

## Select variables of interest
comp_dat <- comp_dat %>% select(LOBBYIST_ID, COMPENSATION_AMOUNT)
contribute_dat <- contribute_dat %>% select(LOBBYIST_ID, AMOUNT)

## Join datasets
lobby_dat  <- inner_join(comp_dat, contribute_dat,
                         on = c("LOBBYIST_ID" = "LOBBYIST_ID"))

## Make lobbyist id a factor for grouping
lobby_dat[, "LOBBYIST_ID"] <- as.factor(lobby_dat[, "LOBBYIST_ID"])

## Clean variable names
colnames(lobby_dat) <- c("lobbyist.id", "comp.amount", "contrib.amount")
pander(head(lobby_dat))
```

lobbyist.id	comp.amount	contrib.amount
8081	52500	1500
8081	52500	250
8081	52500	250
6039	2000	500
6039	2000	100
6039	2000	250

Fine tuning

After some manipulation with `dplyr` I now have a managable dataset with only the variables I care about. However, notice that there are multiple records for each lobbyist. To take care of that I'll use `group_by` and `summarise` to calculate the total values for each lobbyist.

```
# Group by lobbyist, calculate sum for comp and contriubtion
lobby_summary <- lobby_dat %>%
  group_by(lobbyist.id) %>%
  summarise(comp_sum = sum(comp.amount), contrib_sum = sum(contrib.amount))

## View new tibble
pander(head(lobby_summary))
```

lobbyist.id	comp_sum	contrib_sum
5505	169500	2500
5536	7500	1300
5684	1260000	42750
5703	439000	4000
5728	2800	600
5762	548500	28000

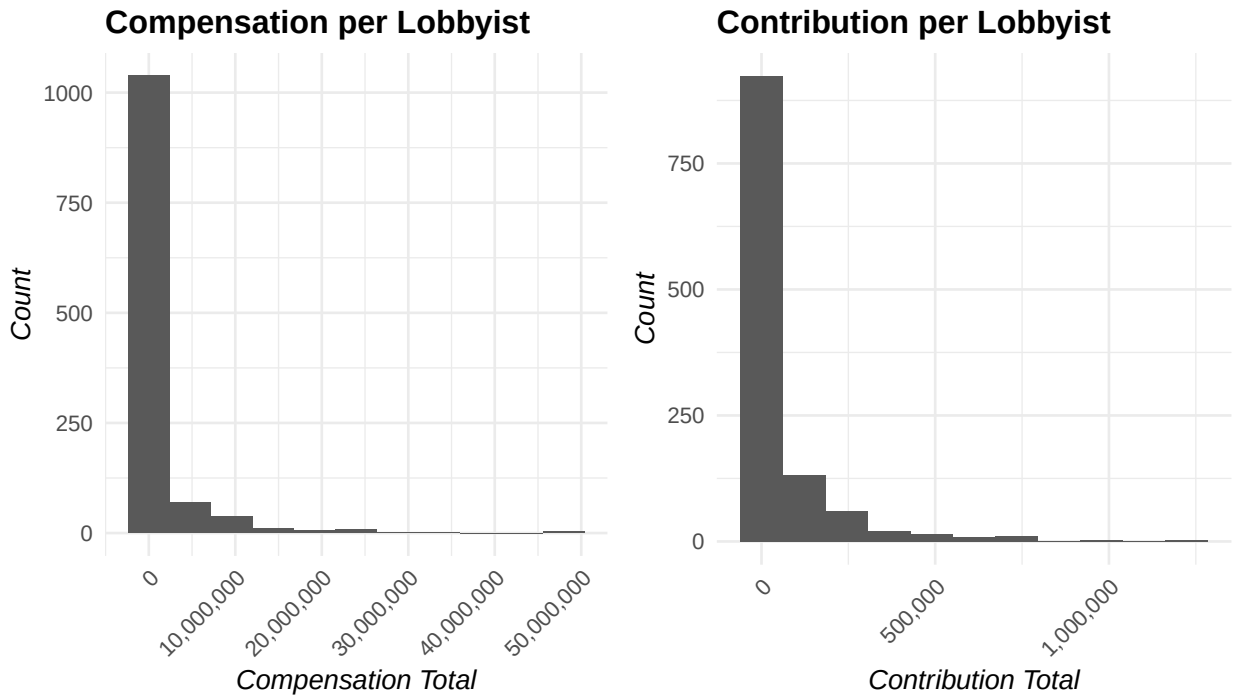
Visualization

```
require(gridExtra)

plot1 <- ggplot(lobby_summary, aes(comp_sum)) +
  geom_histogram(bins = 11) +
  ggtitle("Compensation per Lobbyist") +
  xlab("Compensation Total") + ylab("Count") +
  scale_x_continuous(labels = scales::comma) +
  theme_minimal() +
  theme(axis.title.x = element_text(face = "italic"),
        axis.title.y = element_text(face = "italic"),
        axis.text.x = element_text(angle=45, hjust=1),
        title = element_text(face = "bold"))

plot2 <- ggplot(lobby_summary, aes(contrib_sum)) +
  geom_histogram(bins = 11) +
  ggtitle("Contribution per Lobbyist") +
  xlab("Contribution Total") + ylab("Count") +
  scale_x_continuous(labels = scales::comma) +
  theme_minimal() +
  theme(axis.title.x = element_text(face = "italic"),
        axis.title.y = element_text(face = "italic"),
        axis.text.x = element_text(angle=45, hjust=1),
        title = element_text(face = "bold"))

grid.arrange(plot1, plot2, ncol=2)
```



As shown in the plots above, there is a heavy right skew on both variables. Interestingly, there seems to be a few incredibly well compensated lobbyists. The contributions do not reach such high numbers as it seems to cap around $1,500,000. The big takeaway from these plots is that there are a few incredibly expensive endeavours that companies deem worth the money.

Summary Statistics

```
pander(summarise(lobby_summary, Mean = mean(comp_sum), Median = median(comp_sum),
          Std_dev = sd(comp_sum), IQR = IQR(comp_sum)), caption = "Compensation per lobbyist")
```

Table 3: Compensation per lobbyist

Mean	Median	Std_dev	IQR
1437068	110586	4397683	517688

```
pander(summarise(lobby_summary, Mean = mean(contrib_sum),
                 Median = median(contrib_sum), Std_dev = sd(contrib_sum),
                 IQR = IQR(contrib_sum)), caption = "Contribution per lobbyist")
```

Table 4: Contribution per lobbyist

Mean	Median	Std_dev	IQR
61119	9050	138014	44500

COME BACK HERE AND EXPLAIN

Simple Linear Regression

```
lobby_mod <- lm(data=lobby_summary, contrib_sum ~ comp_sum)
pander(summary(lobby_mod))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27016	2772	9.746	1.235e-21
comp_sum	0.02373	0.0005994	39.59	1.899e-218

Table 6: Fitting linear model: contrib_sum ~ comp_sum

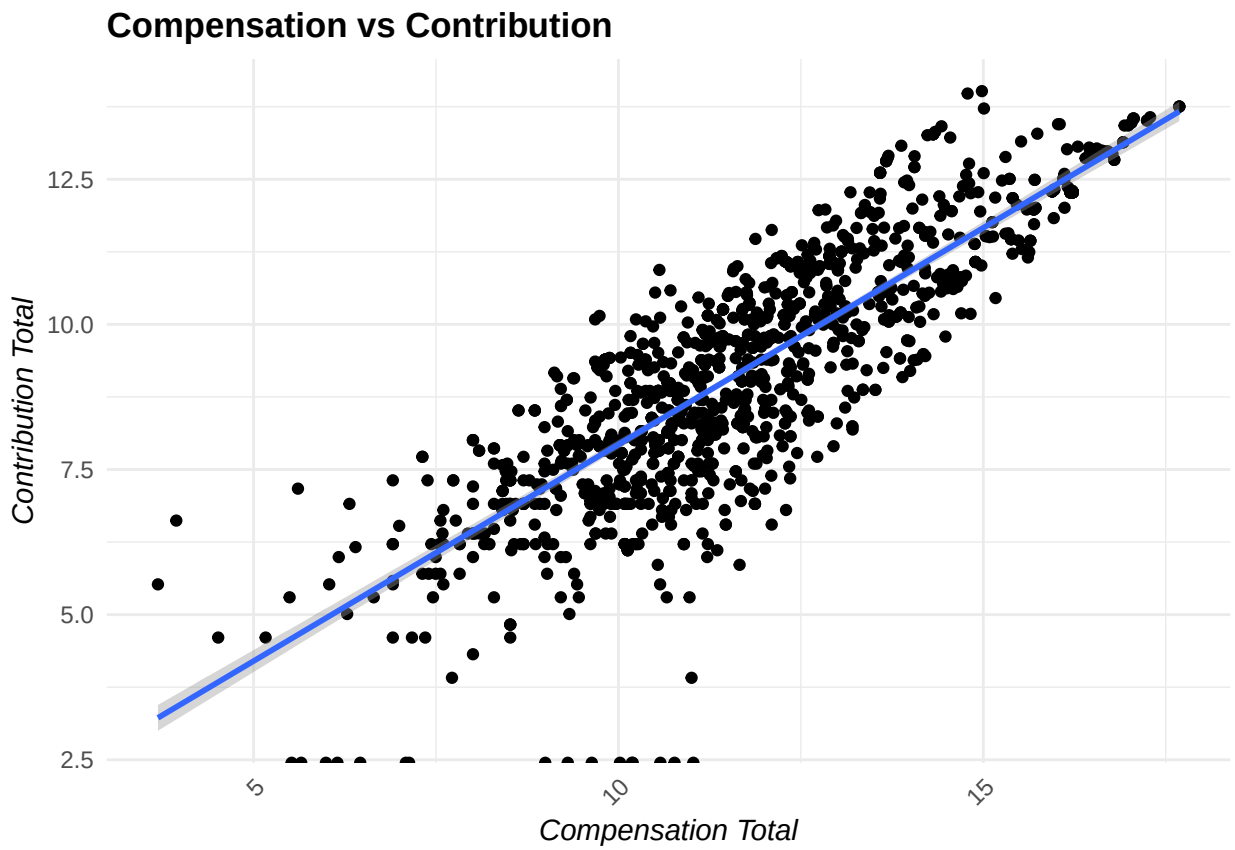
Observations	Residual Std. Error	R^2	Adjusted R^2
1176	90353	0.5718	0.5714

EXPLAIN LINEAR MODEL HERE

Bivariate Plot

```
## Plot linear regression
ggplot(lobby_summary, aes(log(comp_sum), log(contrib_sum))) +
  geom_point() +
  stat_smooth(method="lm") +
  ggtitle("Compensation vs Contribution") +
  xlab("Compensation Total") + ylab("Contribution Total") +
  theme_minimal() +
  theme(axis.title.x = element_text(face = "italic"),
        axis.title.y = element_text(face = "italic"),
        axis.text.x = element_text(angle=45, hjust=1),
        title = element_text(face = "bold"))
```

```
## Warning: Removed 16 rows containing non-finite values (stat_smooth).
```



Bootstrap Confidence Interval