Project 1 SDS 322E

Joshua Bryer

4/1/2022

```
library(kableExtra)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()     masks stats::filter()
## x dplyr::group_rows() masks kableExtra::group_rows()
## x dplyr::lag()        masks stats::lag()
```

```
library(dplyr)
f1<-read_csv("Shark Tank India Dataset.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   brand_name = col_character(),
##   idea = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
f2<-read_csv("ShartankIndiaAllPitches.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   `Episode Number` = col_double(),
##   `Pitch Number` = col_double(),
##   Brand = col_character(),
##   Idea = col_character(),
##   `Investment Amount (In Lakhs INR)` = col_double(),
##   `Debt (In lakhs INR)` = col_double(),
##   Equity = col_character(),
##   Anupam = col_character(),
##   Ashneer = col_character(),
##   Namita = col_character(),
##   Aman = col_character(),
##   Peyush = col_character(),
##   Vineeta = col_character(),
```

```
##   Ghazal = col_character(),
##   Season = col_double()
## )
```

Introduction:

For my project I have elected to use two datasets on Shark Tank India. My datasets names were "ShartankIndiaAllPitches" and "Shark Tank India Dataset". Shark Tank India is a show featuring entrepreneurs pitching their ideas to wealthy investors set in India. They contain variables describing the entrepreneurs pitch, what monetary points they ask for and what they receive. Additionally, they feature points of which investors were present, whether a deal was made, and what episode and season they were when these events occurred. I found these datasets on the website "Kaggle", they were free to download and for the most part ready to use. I find these sorts of things interesting as I am curious what sort of variables are most likely to influence investors to buy in. I expect to see a relationship between the asked amounts for equity, valuation, and investment to be correlated with whether or not there was a deal. Additionally, I expect there to be a relationship between whether a deal was made and whether a certain investor was present.

Tidying

The data does not need any tidying. This is because each variable has its own column, each observation has its own row, and each value has its own cell.

Joining/Merging

In order to join the data set "ShartankIndiaAllPitches" and "Shark Tank India Dataset" I used the left_join() function from dplyr. I combined the dataset by the common ID variables including "brand_name"="Brand". Before joining the datasets, there were 117 observations in each dataset. There were 15 unique id's in the dataset "ShartankIndiaAllPitches" and there were 28 unique id's in the "Shark Tank India Dataset".

d

```
df<-left_join(f1, f2, by=c("brand_name"="Brand","idea"="Idea","deal_amount"="Investment Amount (In Lakhs
glimpse(df)
```

```
## Rows: 117
## Columns: 40
## $ episode_number        <dbl> 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, ...
## $ pitch_number          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...
## $ brand_name            <chr> "BluePine Industries", "Booz scooters", "Hear...
## $ idea                  <chr> "Frozen Momos", "Renting e-bike for mobility ...
## $ deal                  <dbl> 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, ...
## $ pitcher_ask_amount    <dbl> 50, 40, 25, 70, 50, 50, 100, 75, 50, 50, 50, ...
## $ ask_equity            <dbl> 5.00, 15.00, 10.00, 1.00, 5.00, 5.00, 0.25, 4...
## $ ask_valuation         <dbl> 1000.00, 266.67, 250.00, 7000.00, 1000.00, 10...
## $ deal_amount           <dbl> 75, 40, 25, 70, 0, 0, 0, 75, 20, 50, 0, 75, 1...
## $ deal_equity           <dbl> 16.00, 50.00, 30.00, 2.75, 0.00, 0.00, 0.00, ...
## $ deal_valuation        <dbl> 468.75, 80.00, 83.33, 2545.45, 0.00, 0.00, 0....
## $ ashneer_present       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ anupam_present        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ aman_present          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ namita_present        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ vineeta_present       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ peyush_present        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ ghazal_present        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ ashneer_deal          <dbl> 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ anupam_deal           <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, ...
## $ aman_deal             <dbl> 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, ...
## $ namita_deal           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...
```

```
## $ vineeta_deal          <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, ...
## $ peyush_deal           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ ghazal_deal           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ total_sharks_invested <dbl> 3, 2, 2, 1, 0, 0, 0, 1, 1, 2, 0, 2, 2, 0, 0, ...
## $ amount_per_shark      <dbl> 25.0, 20.0, 12.5, 70.0, 0.0, 0.0, 0.0, 75.0, ...
## $ equity_per_shark      <dbl> 5.333333, 25.000000, 15.000000, 2.750000, 0.0...
## $ `Episode Number`      <dbl> 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, NA, 5, 5,...
## $ `Pitch Number`        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, NA, 14...
## $ `Debt (In lakhs INR)` <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 30, 0, 0, 0, NA, 0, 0...
## $ Equity                <chr> "18%", "50%", "30%", "2.75%", "0", "0", "0", ...
## $ Anupam                <chr> "N", "N", "Y", "N", "N", "N", "N", "N", "N", ...
## $ Ashneer               <chr> "Y", "Y", "N", "Y", "N", "N", "N", "N", "N", ...
## $ Namita                <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", ...
## $ Aman                  <chr> "Y", "N", "N", "N", "N", "N", "N", "Y", "N", ...
## $ Peyush                <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", ...
## $ Vineeta               <chr> "Y", "Y", "Y", "N", "N", "N", "N", "N", "Y", ...
## $ Ghazal                <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", ...
## $ Season                <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, NA, 1, 1,...
```

Wrangling

```
#Finding the proportion of deals that Aman was invested in
df%>%
  filter(deal==1)%>%
  filter(aman_deal==1)%>%
  count(aman_deal)%>%
  summarise(aman_prop=n/(sum(df$deal)))
```

```
## # A tibble: 1 x 1
##   aman_prop
##       <dbl>
## 1     0.415
```

```
#How many deals has Namita participated in?
df%>%
  group_by(Namita)%>%
  na.omit()%>%
  count(Namita,deal,na.rm=TRUE)
```

```
## # A tibble: 3 x 4
## # Groups:   Namita [2]
##   Namita  deal na.rm     n
##   <chr>  <dbl> <lgl> <int>
## 1 N          0 TRUE     51
## 2 N          1 TRUE     37
## 3 Y          1 TRUE     14
```

```
#Is there a mean difference in deal amount when Ashneer invests vs. when he does not (continued in visu
plot1<-df%>%
  filter(deal==1)%>%
  select(Ashneer,
         deal_amount)%>%
  arrange(-deal_amount)

head(plot1)
```

```
## # A tibble: 6 x 2
```

ask_valuation	pitcher_ask_amount	deal_amount	deal_residual
600	30	105	75
1000	30	100	70
900	45	100	55
2500	50	100	50

```
##   Ashneer deal_amount
##   <chr>          <dbl>
## 1 <NA>             150
## 2 <NA>             105
## 3 <NA>             100
## 4 <NA>             100
## 5 <NA>             100
## 6 <NA>             100
```

```
#What is the distribution of the difference between the deal amount and what the pitcher asked(continue
plot2<-df%>%
  filter(pitcher_ask_amount<150)%>%
  filter(deal_amount<150)%>%
  filter(ask_valuation<7000)%>%
  mutate(deal_residual=deal_amount-pitcher_ask_amount)%>%
  select(ask_valuation,pitcher_ask_amount,deal_amount,deal_residual)%>%
  arrange(-deal_residual)

kable(head(plot2, 4), booktabs = TRUE) %>%
  kable_styling(font_size = 8)
```

We explored the dataset to answer four different questions. In the first question we looked to find the proportion in which Aman invests and found that they invest 42% of the times. In the second question we found that Namita invested in 14 of the deals. In the last two blocks of code we prepared the data for visualizing, in order to answer whether there is a mean difference in deal amounts when Ashneer invests and when he does not. Lastly, we prepared to answer whether ask_ valuation is correlated with the deal_residual(ask amount-actual amount). **Summary Statistics**

```
#Summary of numerical variable "ask_valuation"
summary(df$ask_valuation)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##      0.01    666.67   1250.00   3852.46   2857.14 120000.00
```

```
#Summary of numerical variable "deal_valuation"
summary(df$deal_valuation)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0   100.0   467.1   500.0  6666.7
```

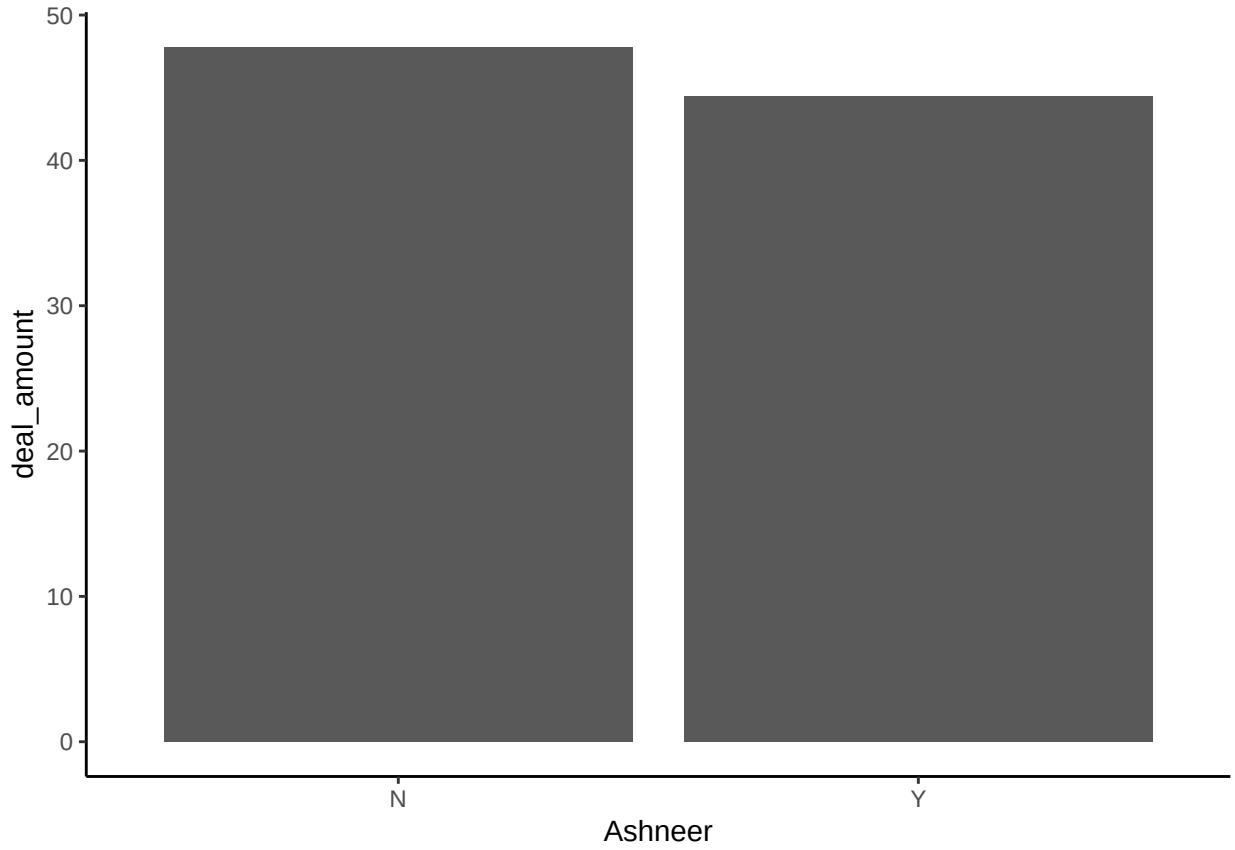
```
#Summary of categorical variable "Ashneer"
table(df$Ashneer)
```

```
##
##  N  Y
## 87 15
```

It can be observed that the range of the ask valuation was 0.01-120,000, the mean was 3852.46, and the median was 1250.00. It can be observed that the range of the ask valuation was 0.0-6666.7 the mean was 467.1, and the median was 100.00. Lastly, the table above shows that on 96 deals Ashneer did not invest and that in 21 they did.

Visualizations

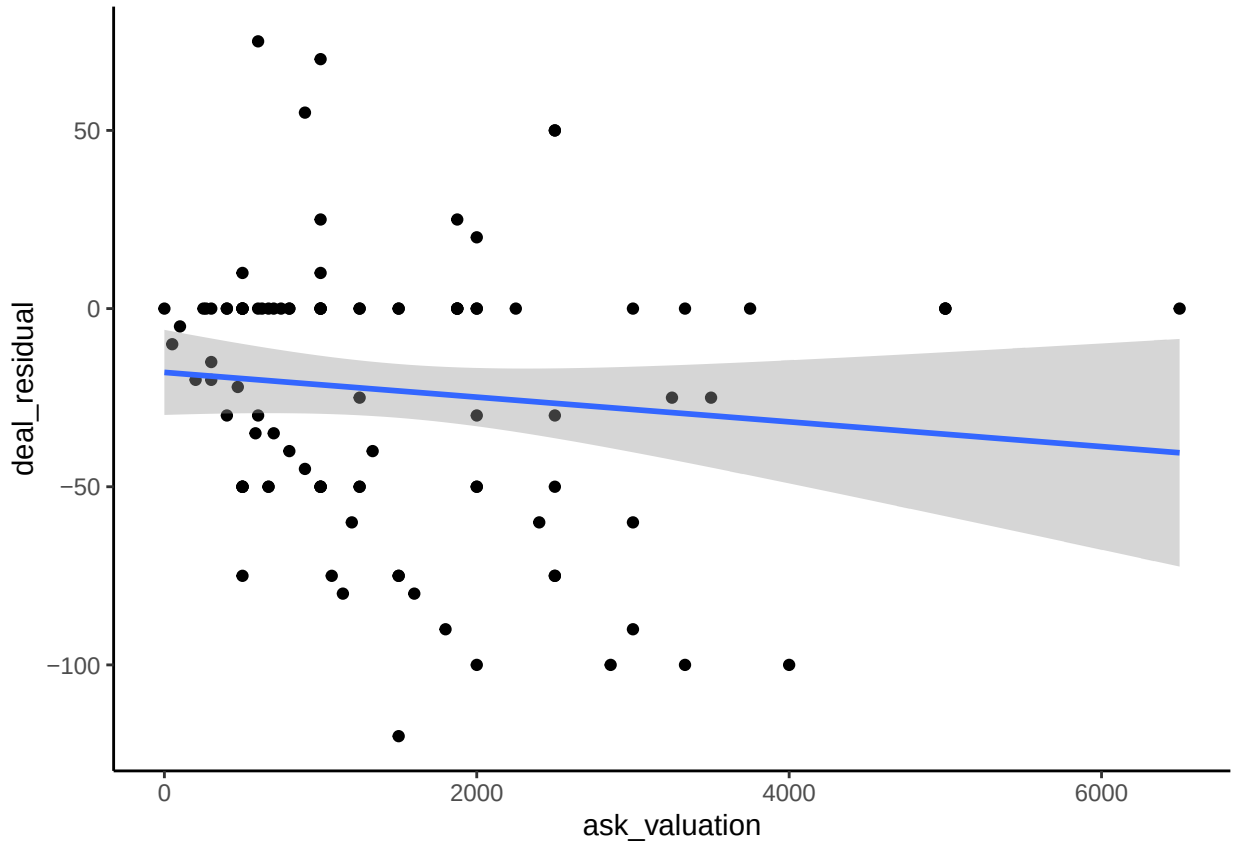
```
#Is there a mean difference in deal amount when Ashneer invests vs. when he does not?
plot1%>%
  na.omit%>%
  ggplot(aes(Ashneer,deal_amount))+
  geom_bar(stat = "summary",fun="mean")+
  theme_classic()
```



In plot one we see that the mean deal amount when Ashneer is invested is greater than the mean of deals where Ashneer is not invested

```
#Does the ask valuation by the pitcher have an effect on the whether the deal is greater or less than t
plot2%>%
  ggplot(aes(ask_valuation,deal_residual,na.rm=T))+
  theme_classic()+
  geom_point()+
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



In this plot we see there is little to no correlation between the ask valuation of the pitcher and whether the deal is greater or less than the ask amount