Problem set #1

John-Henry Pezzuto

Part 1: Write a data section for your assigned data set (5 points)

U.S. patent data

```
library(tidyverse)
library(psych)
library(knitr)
library(kableExtra)
patents <- read_csv("apat63_99.txt")# %>%
#  sample_n(10000, replace = FALSE, weight = NULL, .env = NULL)

theme_set(theme_minimal())
```

Describe how to access data, where it is stored, who curates it. Make sure to use the original source and curator in addition to the NBER site to which I have linked.

To access the patent data, you would go to the http://nber.org/patents/ website and click the links located in the table at the bottom of the page. For the particular data set I am using, you would find the "Patent data, including constructed variables" row within that table and use the link provided under the ASCII CSV column. Alternatively you can simply download the data by clicking here. When using the data you should credit this paper:

> Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights and methodological tools. National Bureau of Economic Research.

The patent data is maintained by National Bureau of Economic Research (NBER). Jean Roth at the NBER seems to be the current contact person for the data, though Bronwyn H. Hall, Adam B. Jaffe and Manuel Trajtenberg are credited in the publication for gathering the data.

Cite other key papers that have used this data.

The original data file has been cited over 3000 times! Some of the more well known papers that have used this data include:

- Acs, Z. J., Anselin, L., & Varga, A. (2002). Patents and innovation counts as measures of regional production of new knowledge. Research Policy, 31(7), 1069–1085. https://doi.org/10.1016/S0048-7333(01)00184-6
- Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 177–187). ACM.
- Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market Value and Patent Citations. The RAND Journal of Economics, 36(1), 16–38.

Describe how the data were collected.

The data includes a list of database on U.S. patents from January 1, 1963 through December 30, 1999 from the United States Patent and Trademark Oce. The researchers took the data offered a step further by creating new indices such as backward and forward citation lags, indices of "originality", which is related to the citations the patent uses, and "generality", how wide the field of patents citing this one is, and self-citations.

Include a table that gives descriptive statistics for at least 8 key variables (you can do more).

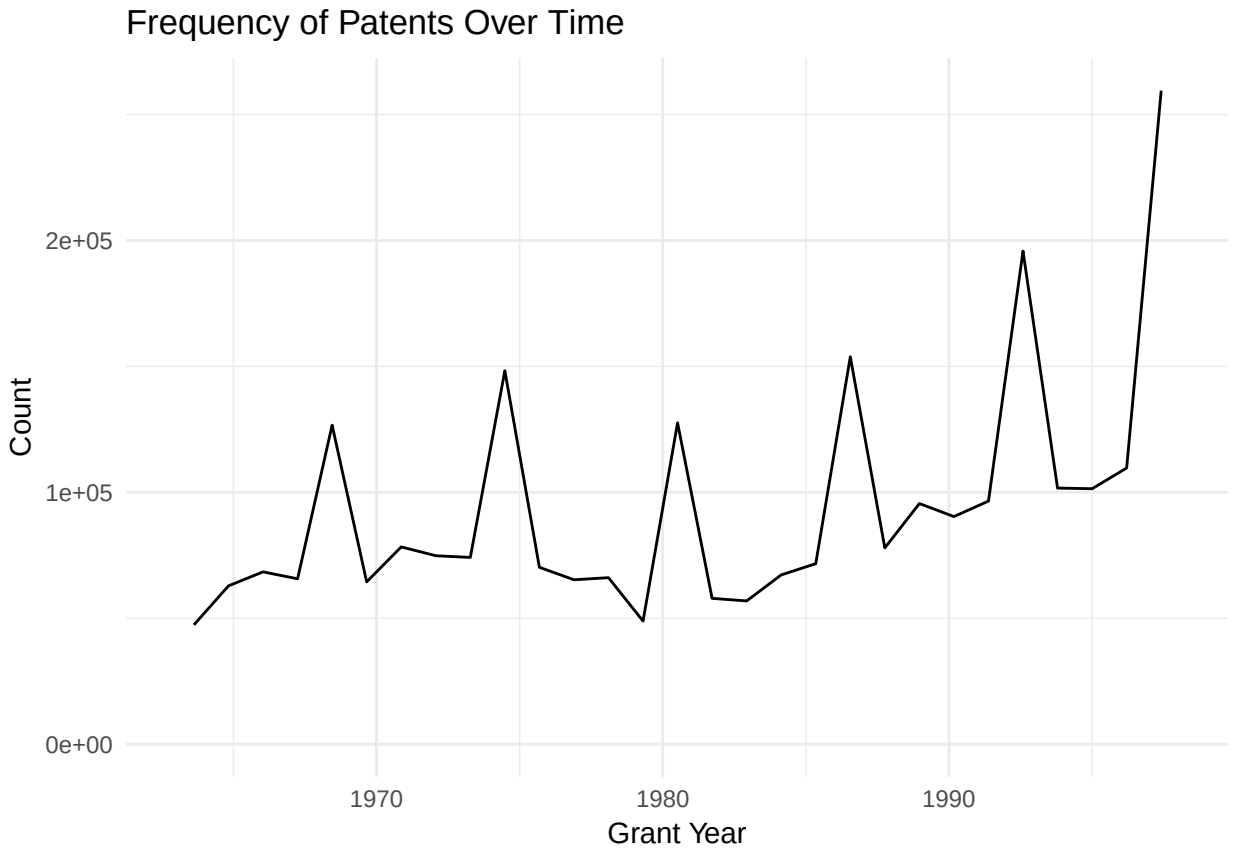
```
patents_clean <- patents %>%
  select(GYEAR, GDATE, APPYEAR, CLAIMS, CMADE, CRECEIVE, GENERAL,
         ORIGINAL, FWDAPLAG, BCKGTLAG, COUNTRY) %>%
  mutate("Grant Year" = as.numeric(GYEAR),
         "Application Year" = as.numeric(APPYEAR),
         "Number of Claims" = as.numeric(CLAIMS),
         "Number of Citations Made" = as.numeric(CMADE),
         "Number of Citations Received" = as.numeric(CRECEIVE),
         "Measure of Generality" = as.numeric(GENERAL),
         "Measure of Originality" = as.numeric(ORIGINAL),
         "Mean Forward Citation Lag" = as.numeric(FWDAPLAG),
         "Mean Backward Citation Lag" = as.numeric(BCKGTLAG),
         "Country" = COUNTRY) %>%
  select(-c(GYEAR, GDATE, APPYEAR, CLAIMS, CMADE, CRECEIVE,
            GENERAL, ORIGINAL, FWDAPLAG, BCKGTLAG, COUNTRY))

describe(patents_clean, skew = FALSE, trim=.1) %>%
  round(3) %>%
  kable(format = "latex") %>%
  kable_styling(latex_options = c("striped", "scale_down"))
```

	vars	n	mean	sd	min	max	range	se
Grant Year	1	2923922	1983.548	10.978	1963	1999.000	36.000	0.006
Application Year	2	2699606	1983.106	10.128	1901	1999.000	98.000	0.006
Number of Claims	3	1984055	12.083	10.268	1	868.000	867.000	0.007
Number of Citations Made	4	2139314	7.720	9.000	0	770.000	770.000	0.006
Number of Citations Received	5	2923922	4.779	7.346	0	779.000	779.000	0.004
Measure of Generality	6	2240348	0.321	0.285	0	0.940	0.940	0.000
Measure of Originality	7	2042151	0.349	0.281	0	0.951	0.951	0.000
Mean Forward Citation Lag	8	2074641	8.306	5.804	0	96.000	96.000	0.004
Mean Backward Citation Lag	9	2088785	14.100	11.769	0	154.000	154.000	0.008
Country*	10	2923922	NaN	NA	Inf	-Inf	-Inf	NA

Include at least one key visualization of the data that exhibits an interesting characteristic.

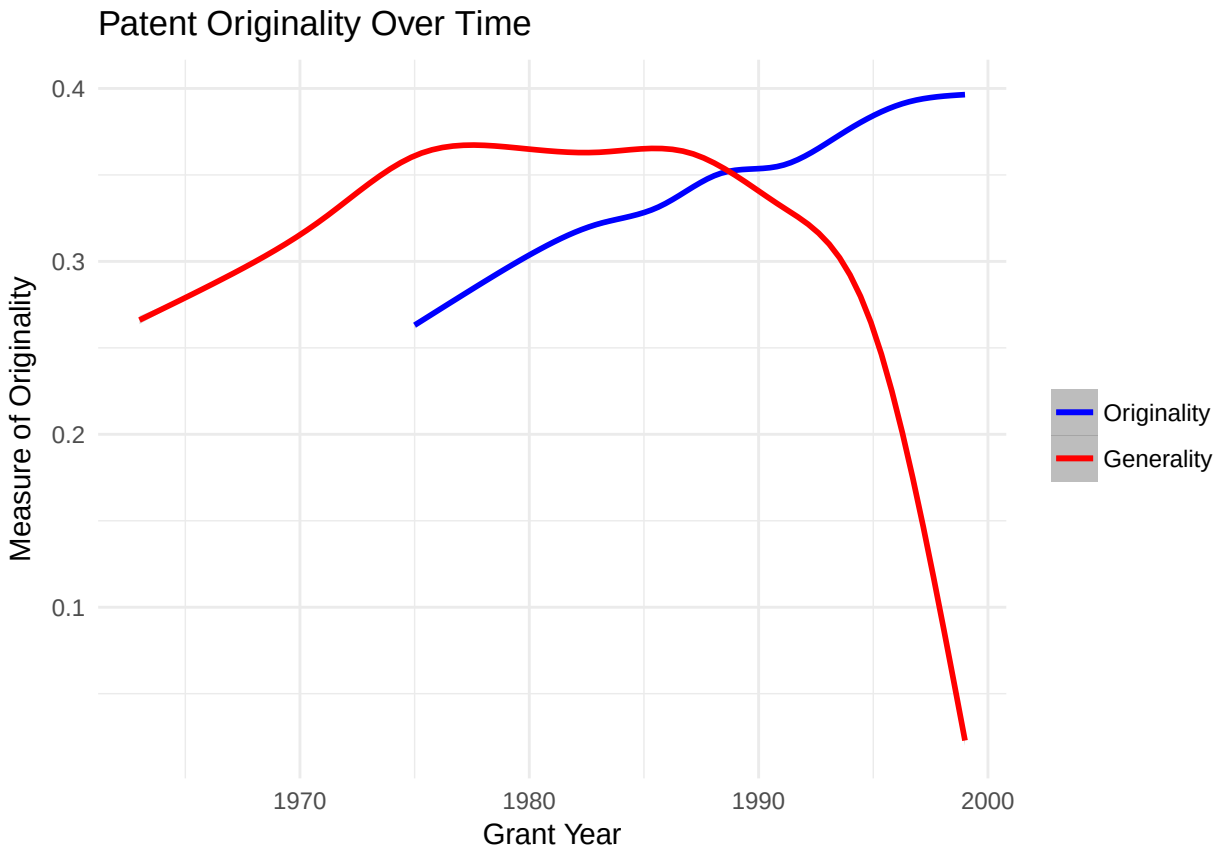
```
patents_clean %>%
  ggplot(aes(`Grant Year`))+
  geom_freqpoly()+
```

```
  scale_x_continuous(limits = c(1963, 1998))+
  labs(title = "Frequency of Patents Over Time",
       y = "Count")
```



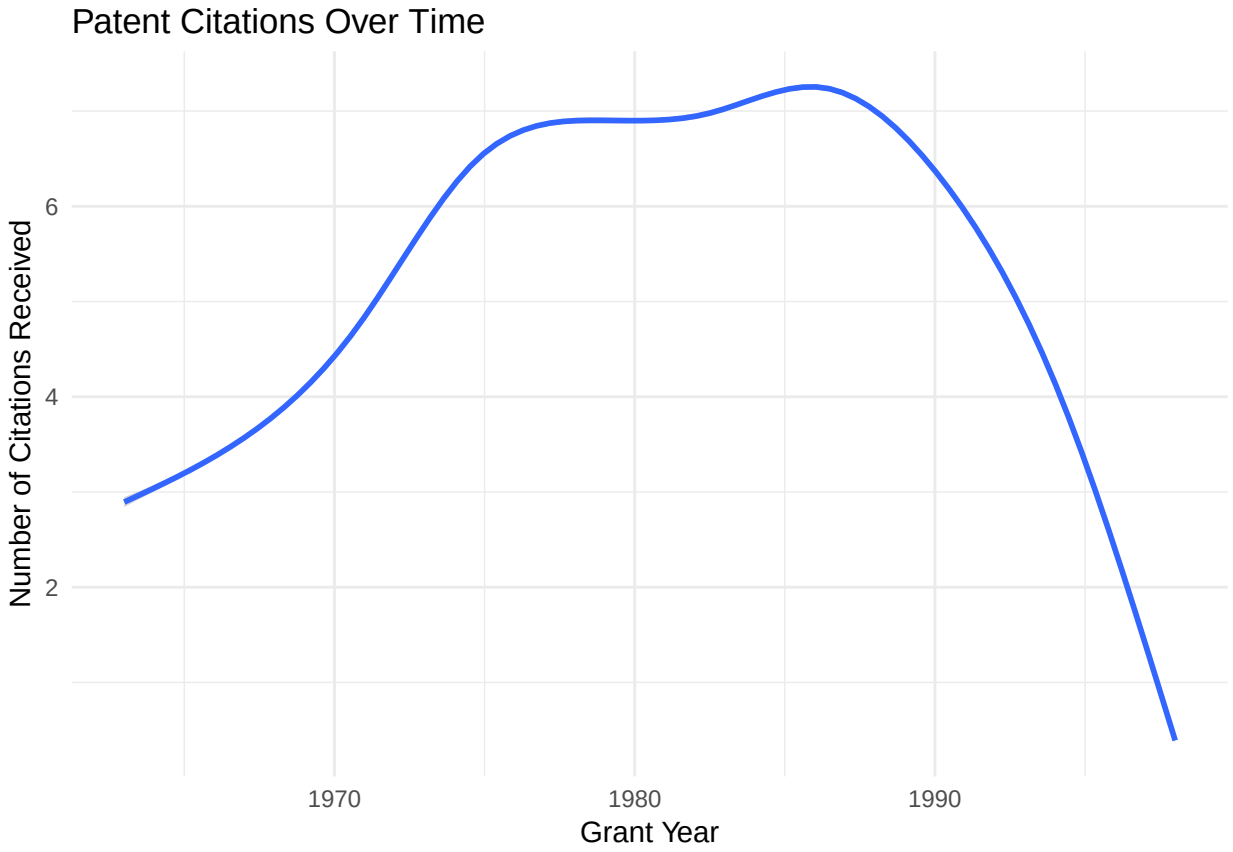
Weird! There seems to spikes every few years!

```
patents_clean %>%
  ggplot(aes(`Grant Year`))+
  geom_smooth(aes(,`Measure of Originality`, color = "blue"))+
  geom_smooth(aes(,`Measure of Generality`, color = "red"))+
  labs(title = "Patent Originality Over Time",
       color='')+
  scale_color_manual(labels = c("Originality", "Generality"), values = c("blue", "red"))
```



This is kind of cool. According to their metric, it seems like patents got more original over time.

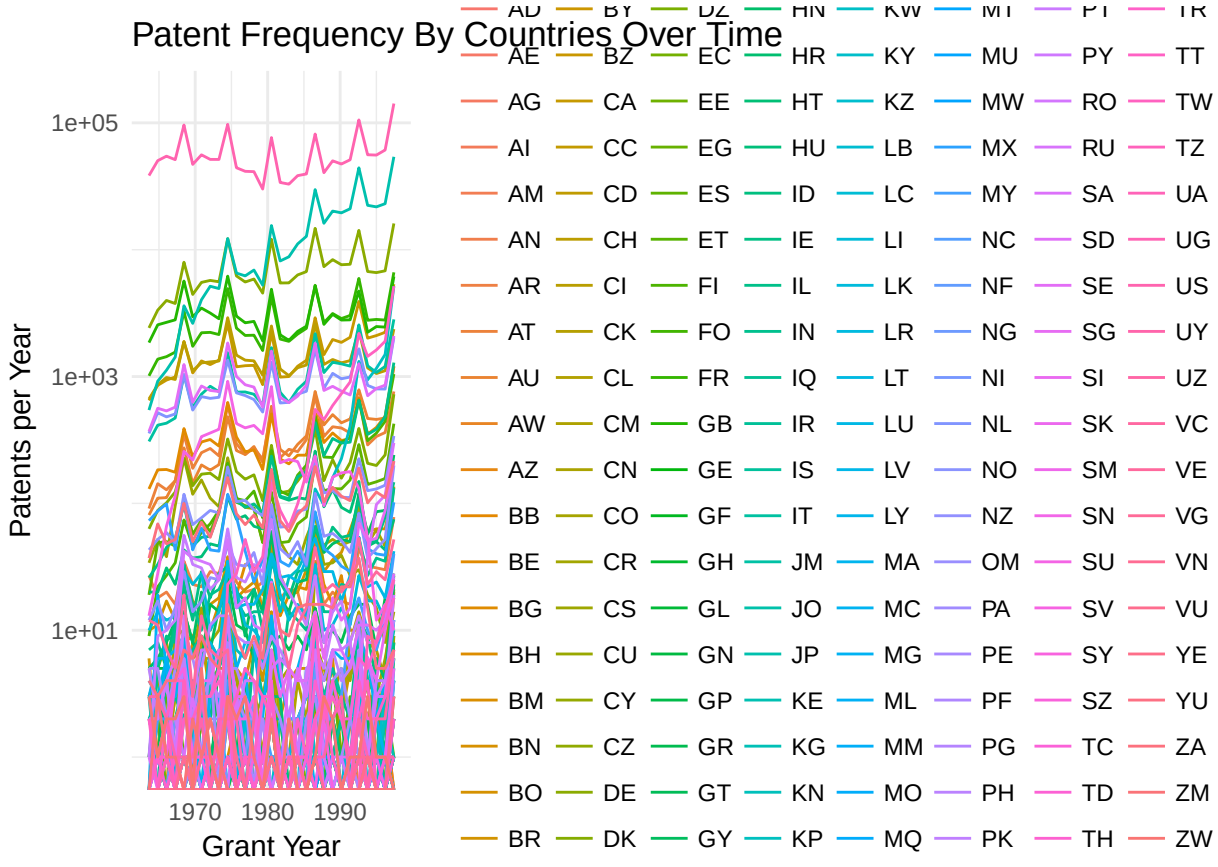
```
patents_clean %>%
  ggplot(aes(`Grant Year`,`Number of Citations Received`))+
  geom_smooth()+
  scale_x_continuous(limits = c(1963, 1998))+
  labs(title = "Patent Citations Over Time")
```



This is a little bit interesting. Intuitively we wouldn't expect the newer patents to receive a lot of citations because they haven't been around for a long time. Perhaps it takes about 10 years for patents to get fully appreciated? Regardless, it seems like the patents from the 80s seemed more highly cited than the patents from the years before.

Show at least one conditional (slice) description of the data (e.g., all variable descriptive statistics by nationality of survey respondent). This can be a table or visualization.

```
patents_clean %>%
  ggplot(aes(`Grant Year`, col = `Country`))+
  geom_freqpoly()+
  scale_y_log10()+
  scale_x_continuous(limits = c(1963, 1998))+
  labs(title = "Patent Frequency By Countries Over Time",
       y = "Patents per Year")
```



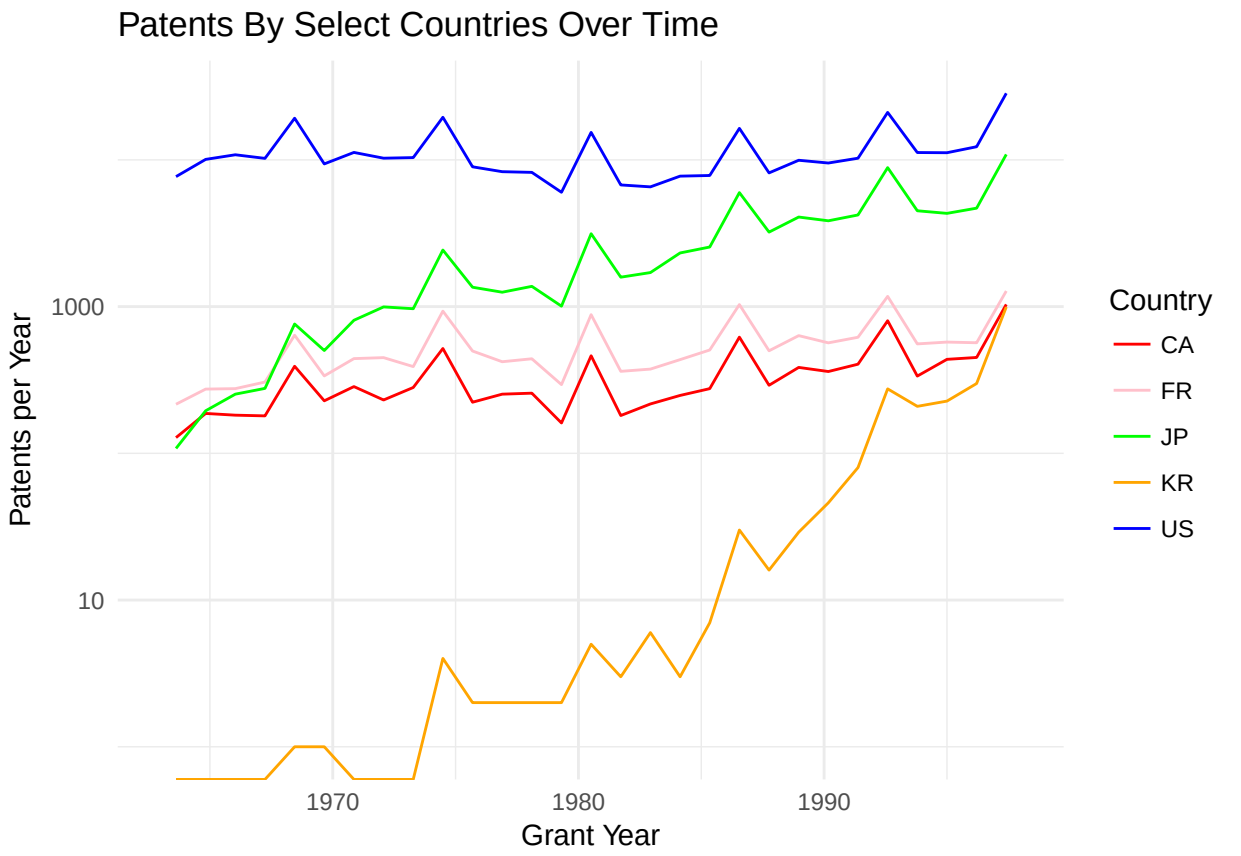
I tried to visualize patents by countries. There are too many countries for this to be meaningful. I am going to try again, after choosing a select few countries.

```
patents_clean %>%
  group_by(Country) %>%
  summarise(n()) %>%
  arrange(desc(`n()`)) %>%
  head(10) %>%
  kable()
```

Country	n()
US	1784989
JP	421441
DE	221095
GB	98012
FR	85398
CA	53872
CH	43313
IT	32433
SE	28286
NL	26687

```
patents_clean %>%
  filter(Country == c("US","JP", "KR", "FR", "CA")) %>%
  ggplot(aes(`Grant Year`, col = `Country`))+
```

```
  geom_freqpoly()+
  scale_y_log10()+
  scale_x_continuous(limits = c(1963, 1998))+
  scale_color_manual(values = c("red", "pink", "green", "orange", "blue"))+
  labs(title = "Patents By Select Countries Over Time",
       y = "Patents per Year")
```



We can now see that US dominates in patents every year, with Japan further behind (the log y scale makes the countries appear much closer than the above table showed). Interestingly during Korea's economic boom in the 90s they seemed about matched with France and Canada!

Part 2: Critique a computational research paper (5 points)

Anagol, S., Cole, S., & Sarkar, S. (2017). Understanding the advice of commissions-motivated agents: Evidence from the Indian life insurance market. Review of Economics and Statistics, 99(1), 1-15.

State the research question of your assigned paper.

The main research question in the paper is **Do commission-motivated agents in India provide quality recommendations to uninformed customers?** The paper also looks at what the impact of the disclosure of fees for the unit linked insurance policy, and how knowledgeable are the agents about the advice the give.

What data did the paper use?

The data the researchers used came from the field experiments they conducted in India. The researchers trained individuals and then sent them to life insurance agents. The individuals were undecided between two plans, one of which was better in absolute terms. The initial beliefs of the insurance seekers were varied in a way that could show if the agents were more responsive to customer's needs, more responsive to customer's incorrect beliefs, or more responsive to commission incentives.

What theory did the paper reference in order to interpret the data? (Note: it is possible that the paper has no reference to theory.)

This paper compares two theories comparing the incentives of agents. The first theory posits that the agents give bad advice to the customers because they have an incentive to suggest the customers use high commission options. The counter theory posits that because the agents are driven by reputation concerns, that they will give customers good advice.

Was your assigned paper a descriptive study, an identification exercise, a numerical solution to system of equations study, or some combination of the three? (These are the three classifications we discussed in class.)

My assigned paper was a combination of a descriptive study and a identification exercise, although it was primarily the latter. The authors collected their own data, and a fair amount of time describing it, and then spend the rest of the paper trying to distinguish the mechanism discerning the quality of the advice, the knowledge of the parties involved, and when the agents are motivated by commissions.

What computational methods did this paper use to answer the research question? What was their result or answer to the question?

This paper did not seem computationally intensive. The researchers used regression.

The researchers found that agents pushed whole insurance projects even when term insurance seemed like a better fit for the customer based on their scripts. They did find however that agents respond to customer's beliefs even when they were incorrect – even when it would be in their best interest to correct them to sell a higher commission product.

Think of yourself as an academic referee. Give two suggestions to the author(s) of your assigned paper of things the authors might do to improve their results or strengthen their evidence for the answer to the question.

1) I would be concerned about the external validity of the findings. Because the study was done in India, it is possible that the results do not extend to other countries. If the authors could recreate their study in other countries I would recommend the authors try to recreate their experiment in multiple countries. It is unclear that the models the authors are comparing would be the same in developed countries.
2) How can the authors be sure there was no communication between agents? As the experiment took place over a fair amount of time, it is possible that if one agent found out, they could have told other agents that there was something fishy going on. In essence, if the agents were "spooked", they might not act like themselves. I would suggest that the authors try to determine the social networks of the agents they are interested in studying to minimize overlap, to prevent this from happening.