Project Part I

9/26/2019

– Background information

In 2016, Russia's Internet Research Agency (IRA) conducted a well-funded social media operation to influence the U.S. presidential election. The IRA exploited several social media platforms, one of which was Facebook. The House Permanent Select Committee on Intelligence (HPSC) uncovered these facts during its open hearings with social media companies in November 2017. In cooperation with the Committee, Facebook conducted an internal investigation and identified 3516 advertisements purchased by the IRA. These ads were viewed by 11.4 million American users. The owner of russian-ad-explorer.github.io compiled the pdf descriptions of these ads and grouped them to make this data set.

– Data Description

Those advertisements comprise the rows of this data set. Each column represents a characteristic of those ads. The compiler didn't define every column, so I'll have to infer some meanings. I'll be studying the following columns: - Clicks: the number of times an ad was clicked. - Impressions: the number of times an ad appeared on a person's screen. - Creation Date: Date that the ad was created. - End Date: Date that the ad was discontinued. - Location: the city, state, country, or region targeted by the ad. - Adspend: Currency amount that the IRA-linked page spent to place an ad.

```
ads = read.csv('FacebookAds.csv')
dim(ads)
```

```
## [1] 3516   25
```

The dataset contains 3516 rows, one for each advertisement.

– Data Representation and Collection Method

Based on government files (reference link 2), all IRA-linked advertisements that Facebook could find have been included in this data set. In this way, the set is not a sample but may represent the population of russian linked ads placed on facebook between June 2015 and August 2017. The HPSCI memo (link 2) states that these ads were purchased by the Russian IRA, so Facebook must have identified ads through these transactions.

– Potential Issues

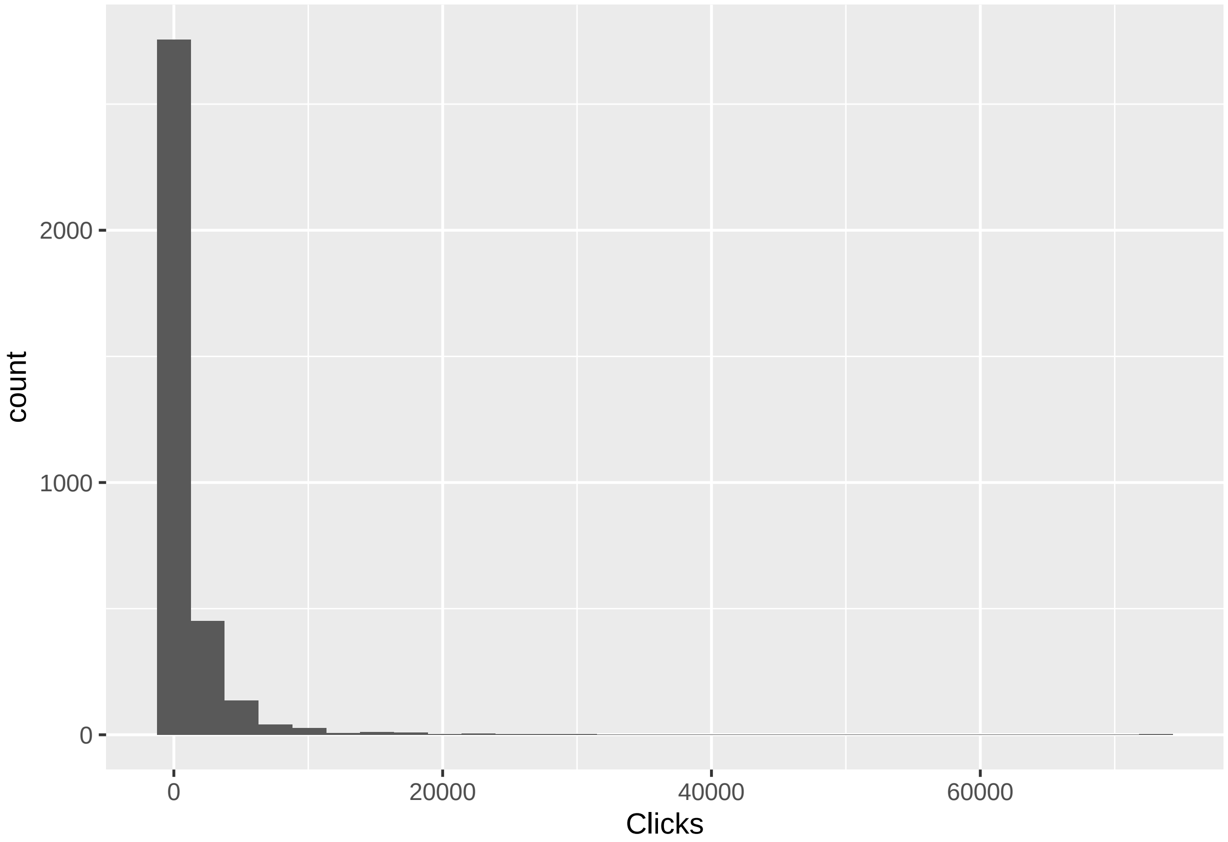
I may have issues parsing the location strings, since elements of this column contain many locations and area types. Still, I hope to identify a few locations that were heavily targeted. I may also have to convert from Rubles to Dollars and vice versa.

I'm interested in exploring whether I can predict number of clicks or interactions with advertisement expense, geographic location (categorical), and ad air time, or whether a parameter of number of interactions or clicks is significantly different for two different groups. I'll implement cross validation to identify the best model. Therefore, I will need the following plots:

```
click.hist = ggplot(ads, aes(x=Clicks))+geom_histogram()
click.hist
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 60 rows containing non-finite values (stat_bin).
```

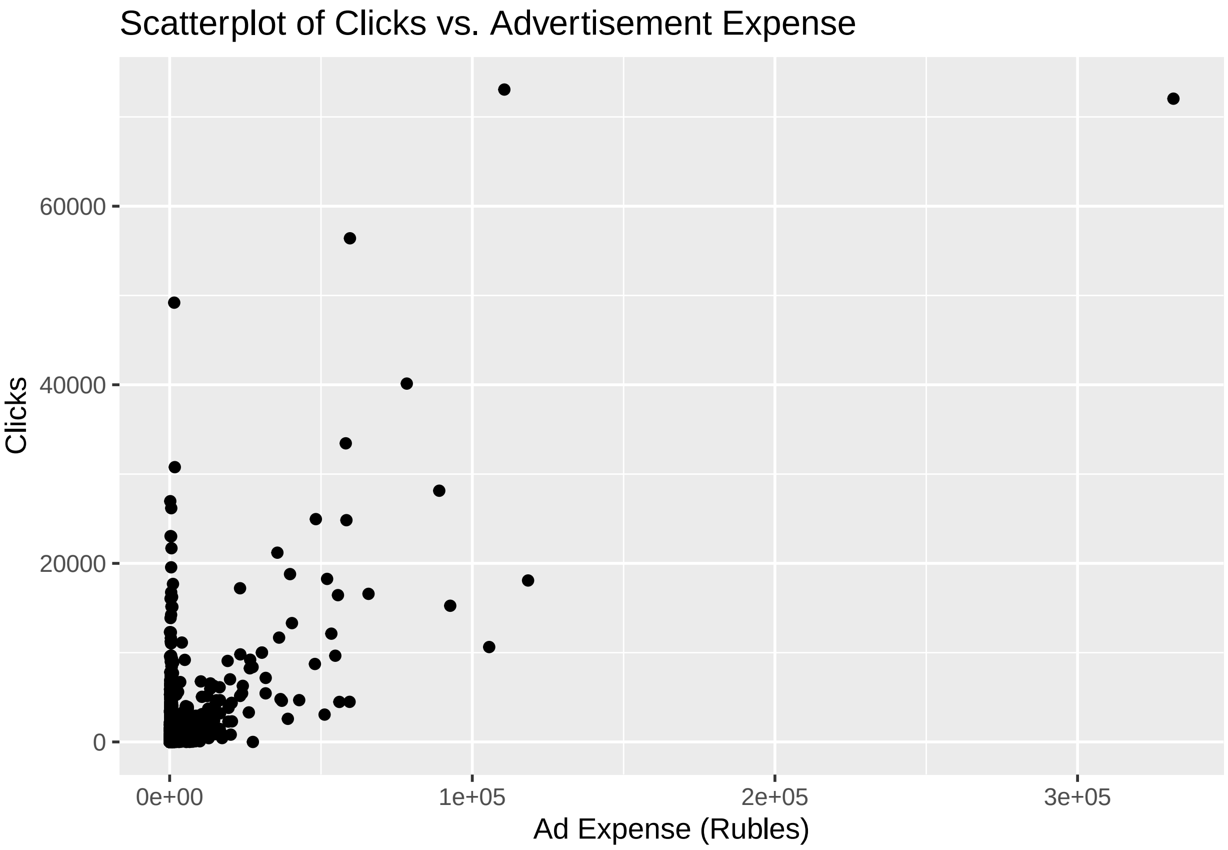


The histogram of clicks is not normal, so I'll probably have to conduct a box-cox transformation of the response if I want to avoid a violation of the regression normality assumption and have randomly distributed residuals.

For Clicks vs. each Independent Variable:

```
#units? Rubles only- USD rows have no dollar counts in the AdSpend col
#unique(ads$AdSpendCurrency)
click.spend = ggplot(ads, aes(x=AdSpend, y=Clicks))+geom_point()+labs(x="Ad Expense (Rub
click.spend
```

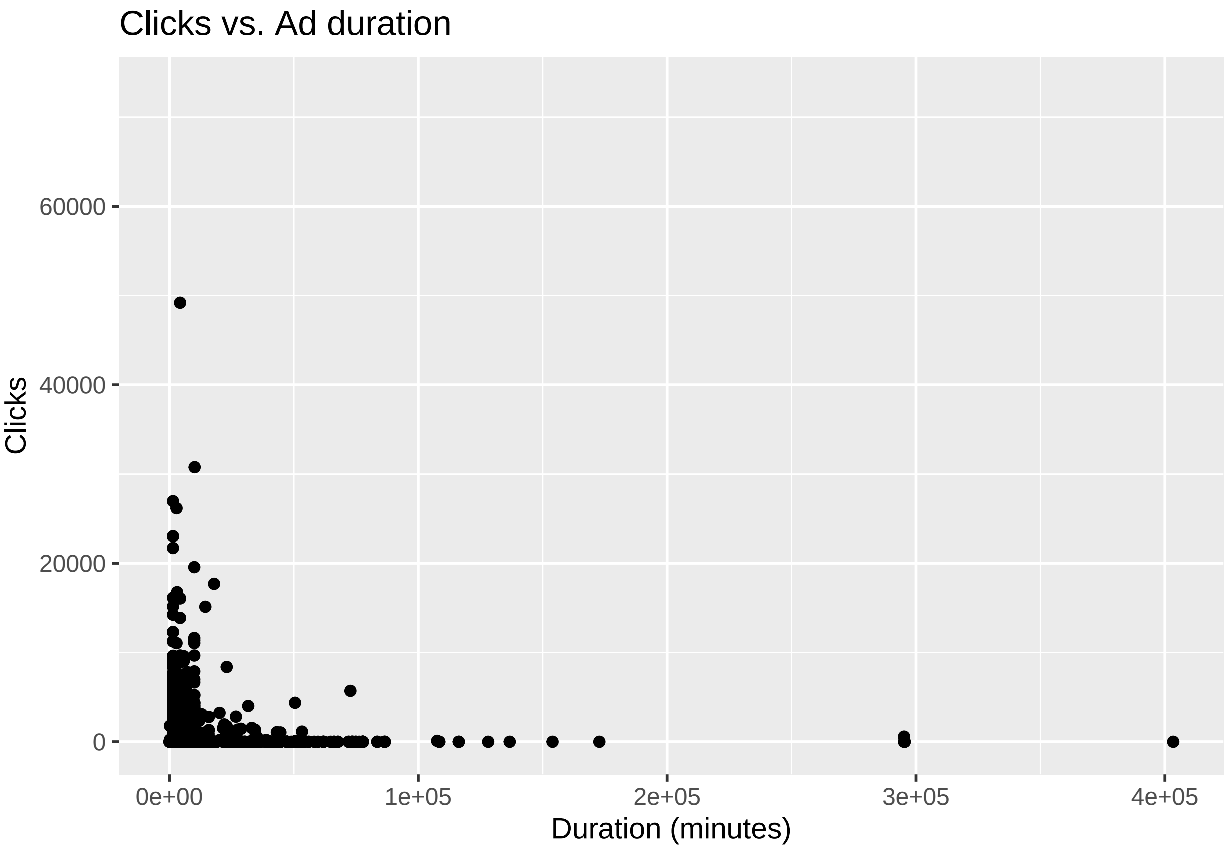
```
## Warning: Removed 1049 rows containing missing values (geom_point).
```



```
create = mdy_hms(ads$CreationDate)
end = mdy_hms(ads$EndDate)
duration = end-create
ads$Duration = abs(duration)
duration.clicks = ggplot(ads, aes(x=Duration, y=Clicks))+geom_point()+labs(x="Duration (
duration.clicks
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to

## Warning: Removed 1264 rows containing missing values (geom_point).
```



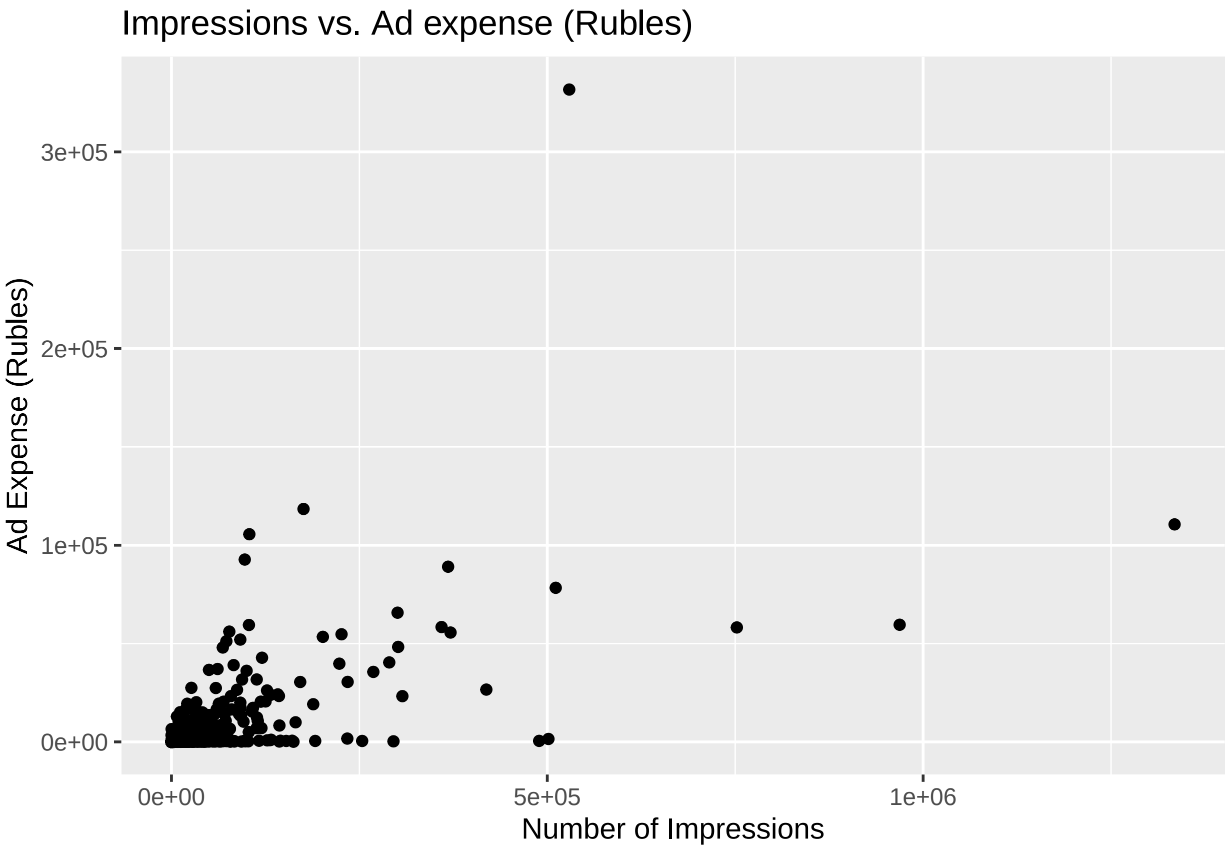
The plot of Duration vs. Clicks exhibits a promising possibly quadratic trend. I may want to include a squared Duration term in my model.

The data cleaning for the xlicks vs location boxplot is a little more involved, so I'll save it for the next submission.

For Impressions vs. each IV:

```
int.spend = ggplot(ads, aes(x=Impressions, y=AdSpend))+geom_point() +labs(title='Impress
int.spend
```

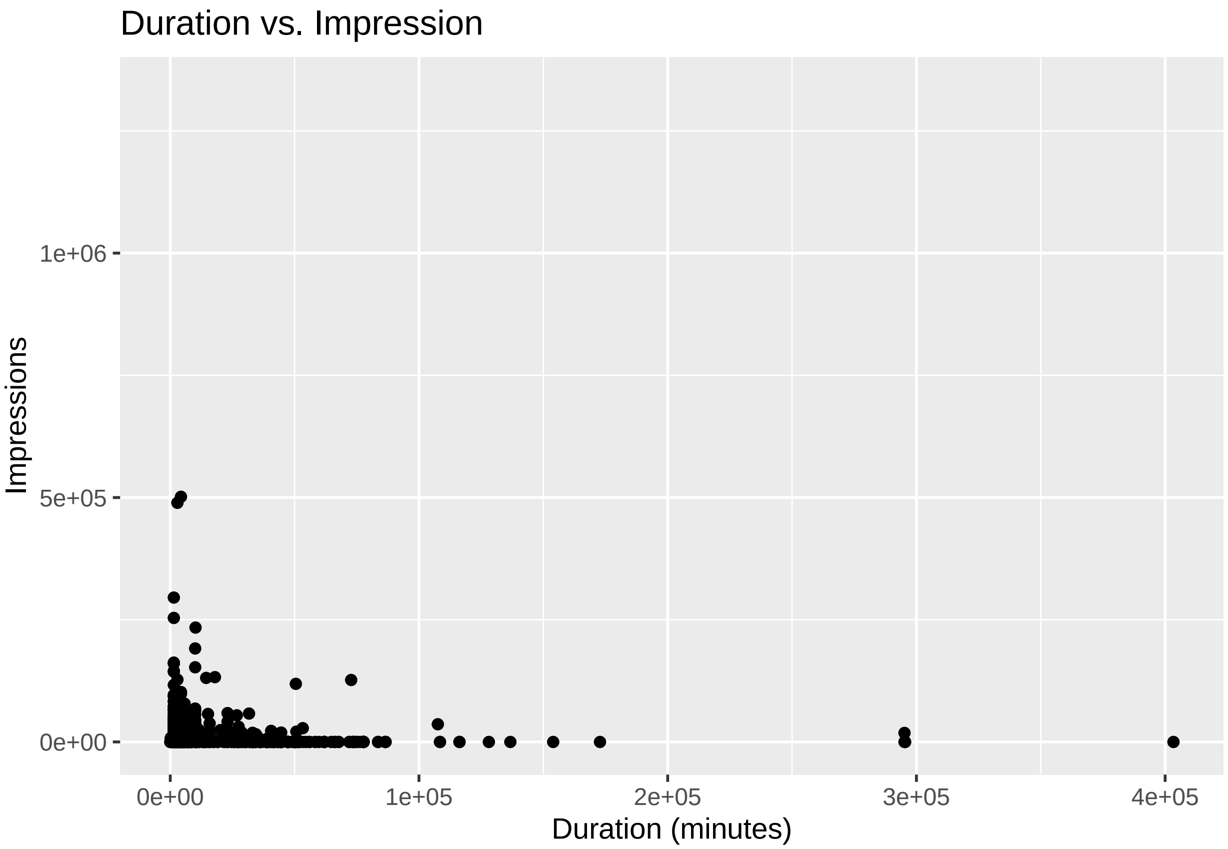
```
## Warning: Removed 1048 rows containing missing values (geom_point).
```



```
duration.impression = ggplot(ads, aes(x=Duration, y=Impressions))+geom_point()+labs(x="D
duration.impression
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to
```

```
## Warning: Removed 1272 rows containing missing values (geom_point).
```



The plot of Duration vs. Impressions exhibits a another potentially quadratic trend. I may want to include a squared Duration term in my model for predicting Impressions. Missing observations may pose a problem here though.

Numerical summaries

```
numsum = c(mean(ads$Clicks, na.rm=TRUE), median(ads$Clicks, na.rm=TRUE))
c(paste("mean=", numsum[1]), paste("median=", numsum[2]))
```

```
## [1] "mean= 1079.56336805556" "median= 75"
```

The advertisements were each clicked 1079.563 times on average. However, this number may be influenced by outliers considering the histogram of clicks. Therefore, a median may be a better representation of the typical ad. The median number of clicks on an advertisement is 75, which is way lower than the mean. This indicates that the distribution of clicks is heavily skewed right. This may be problematic for the regression.

References

1. https://www.kaggle.com/paultimothymooney/russian-political-influence-campaigns.
2. <intelligence.house.gov/uploadedfiles/hpsci_minority_exhibits_memo_11.1.17.pdf>.
3. Detailed data descriptions: https://russian-ad-explorer.github.io/about