|  |
| --- |
| BRICS 2018 Skills Challenge  Nabeel and Louise Day 1 Challenge |

|  |  |
| --- | --- |
| Which factors influence commuters’ behaviour?01Fuel Price, CPI and the Unemployment Rate affect the cost of transport02Are people more likely to share a ride if transportation costs increase?03 | Abstract This solution will aim to prove that certain environmental factors influence the number of passengers that commute together. The hypothesis tests whether changes in fuel price, unemployment rate, inflation or temperature are not only correlated, but also causes changes in commuters sharing transport. This solution makes use of the Walmart and Taxi Fare datasets. We are omitting any rows where the passenger count or the fare amount is equal to or below one. Without further context around the data, no conclusion can be made and hence, these data points are not of interest to the model. The loss of these subsets and the influence on the models, are outside of the scope of this solution. The Walmart dataset does not include any geographical data. The Taxi Fare dataset are limited to New York City. Assuming that all Walmart stores are in the US, the solution joins the two datasets as if they are geographically similar. |

# Selected Dataset & Data Characteristics

Below are two tables that describe each data sets’ characteristics:

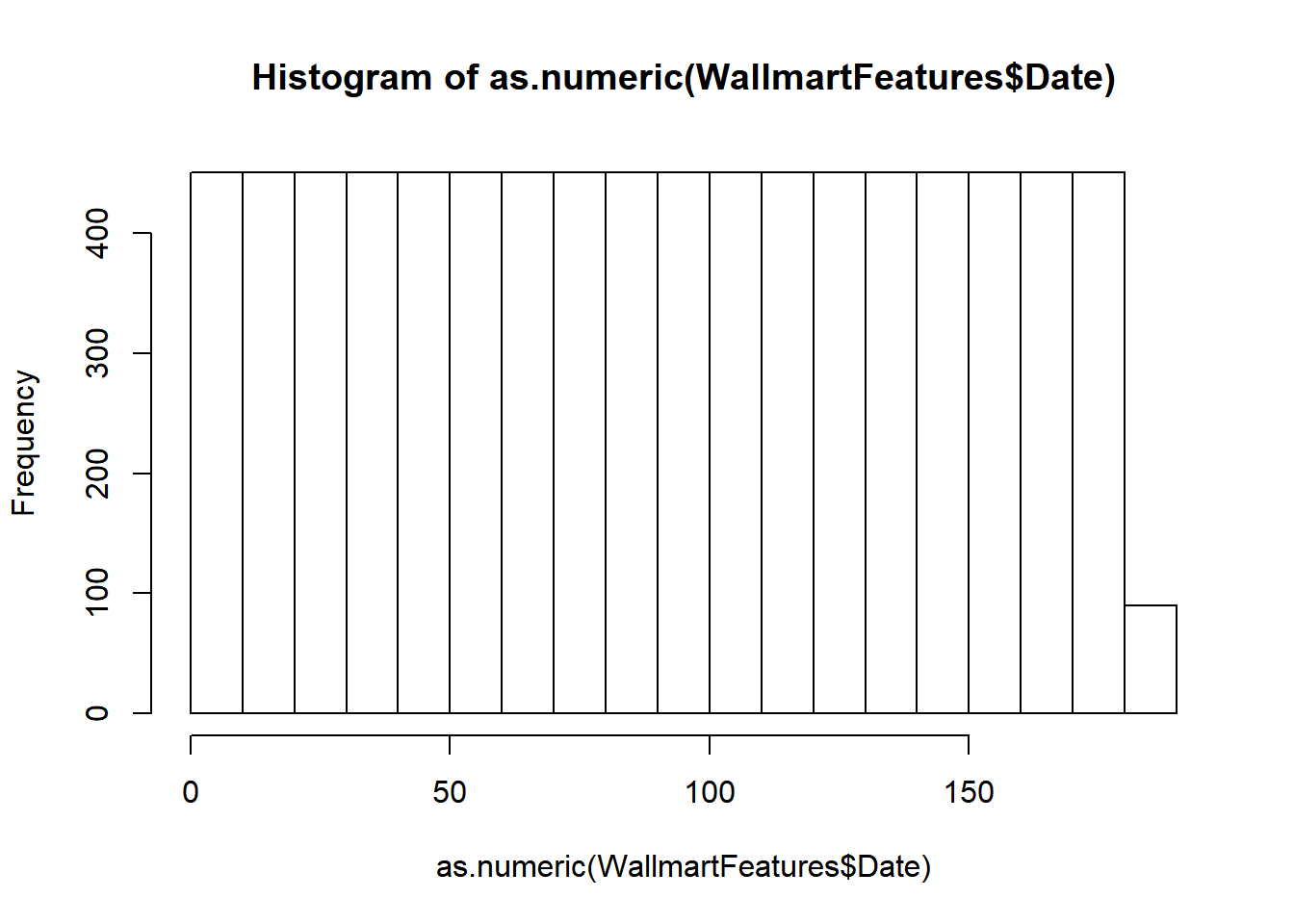
|  |  |  |
| --- | --- | --- |
| Walmart Dataset | | |
| Column Name | Data Type | Comments |
| Store Id | Integer – key | Not used |
| CPI | Decimal | Used as inflation data for New York City |
| Fuel Price | Decimal | Used as fuel price data for New York City |
| Unemployment Rate | Decimal | Used as unemployment data for New York City |
| Temperature | Decimal | Used as temperature data for New York City |
| Date | Date time |  |
| Markdown1 | Unknown | Not used |
| Markdown2 | Unknown | Not used |
| Markdown3 | Unknown | Not used |
| Markdown4 | Unknown | Not used |
| Markdown5 | Unknown | Not used |
| Is Holiday | Boolean |  |

|  |  |  |
| --- | --- | --- |
| Taxi Fare Dataset | | |
| Column Name | Data Type | Comments |
| Key | String – Key | Not used |
| Number of Passengers | Integer | Only observations greater than 1 is used |
| Fare Amount | Decimal | Only observations greater than 1 is used |
| Longitude of pick up location | Decimal |  |
| Latitude of pick up location | Decimal |  |
| Longitude of drop off location | Decimal |  |
| Latitude of drop off location | Decimal |  |
| Pick up date and time | Date time |  |

# Data exploration & Observations

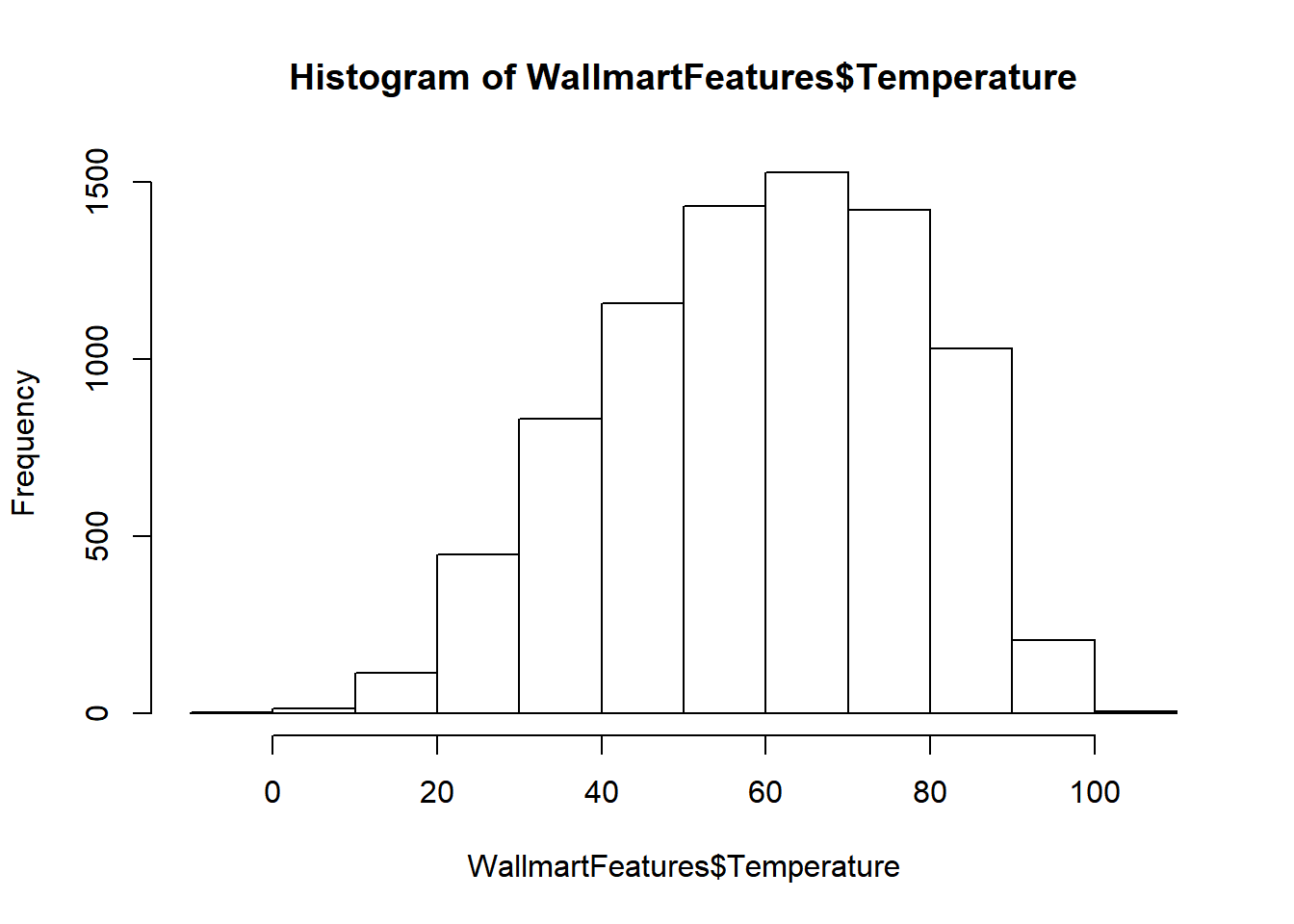
## Walmart Dataset

### Date Distribution

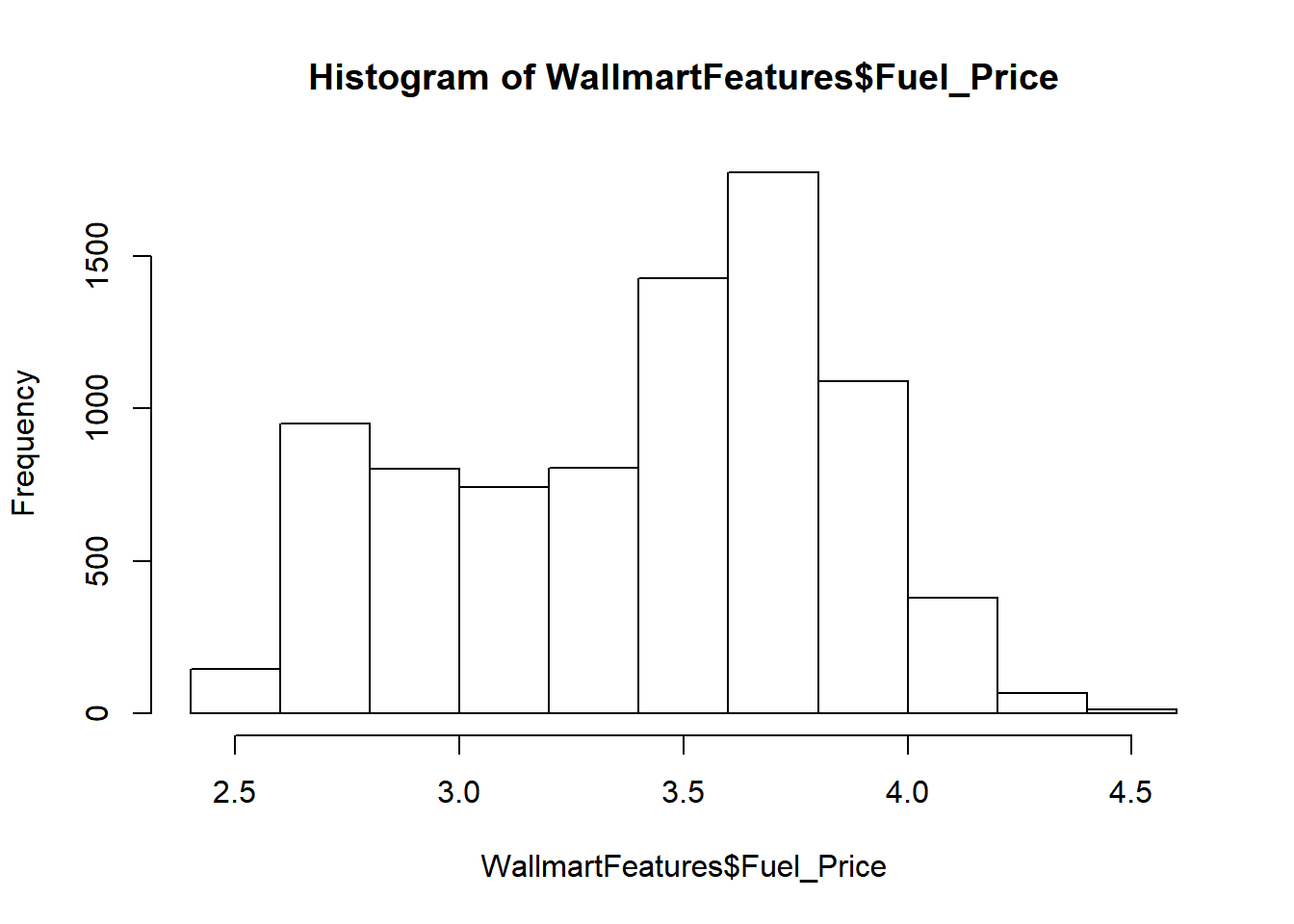


Equal data points exist for the date range of 2010/02/05 to 2013/07/26

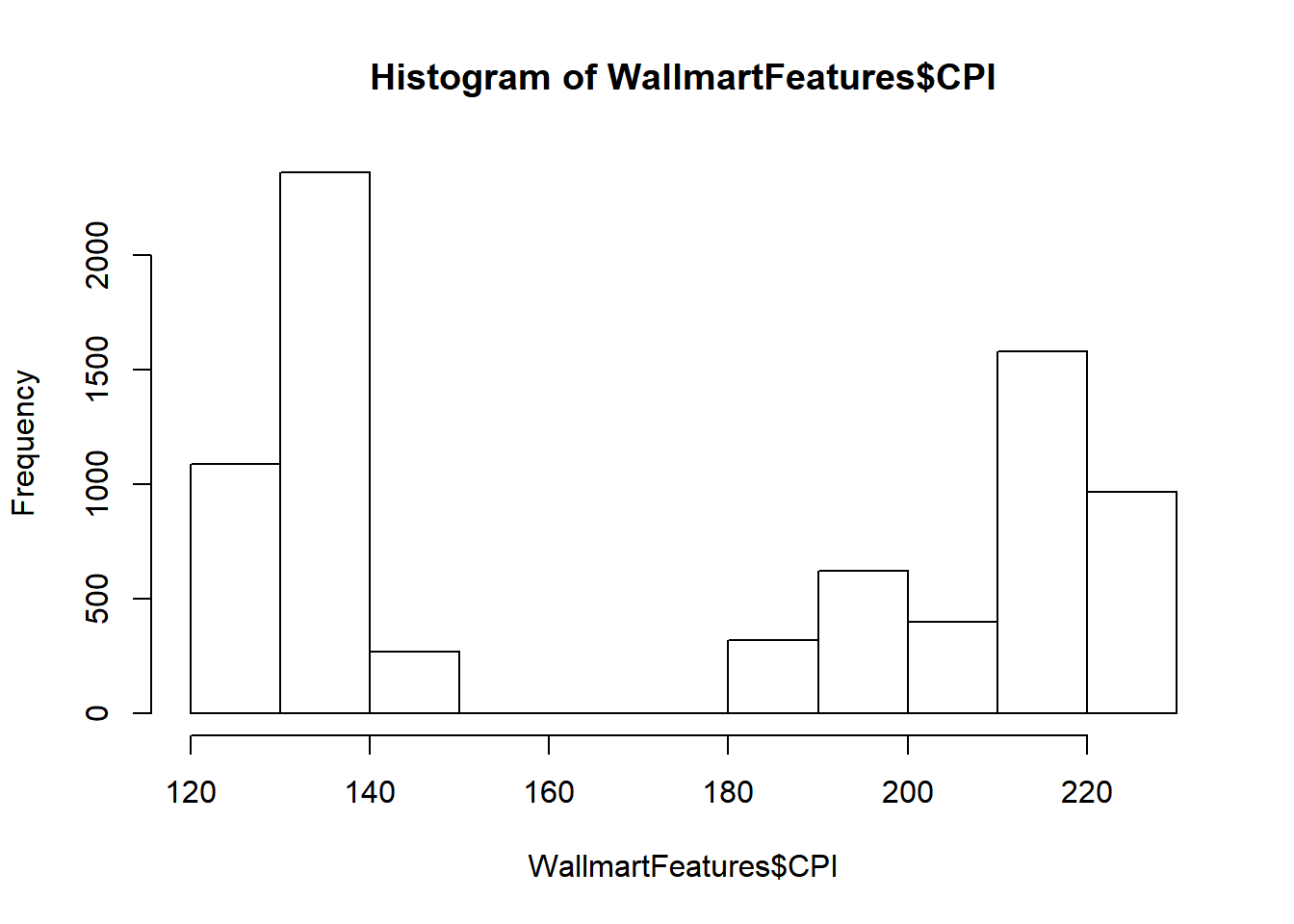
### Temperature Distribution



### Fuel Price Distribution

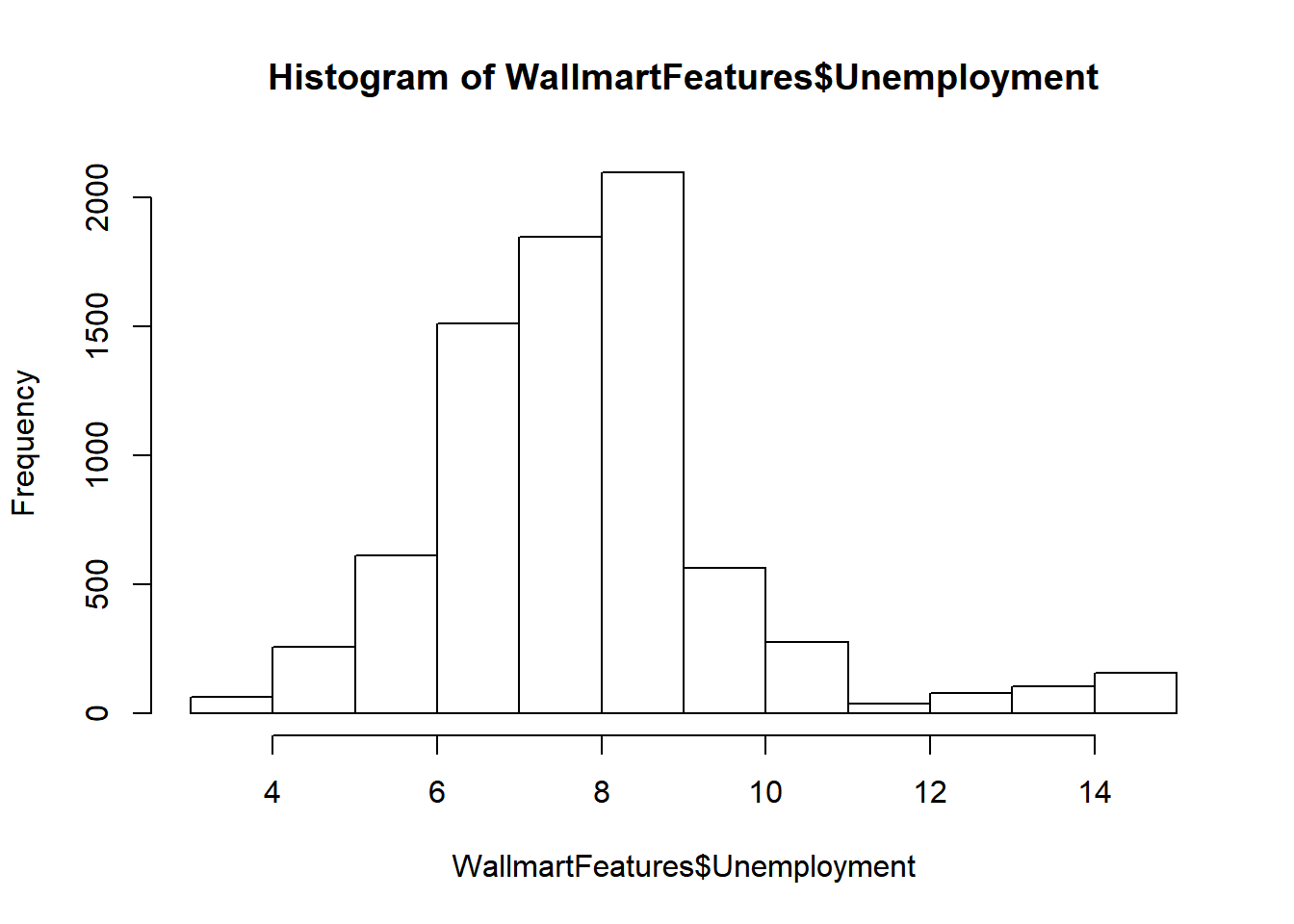


### CPI Distribution

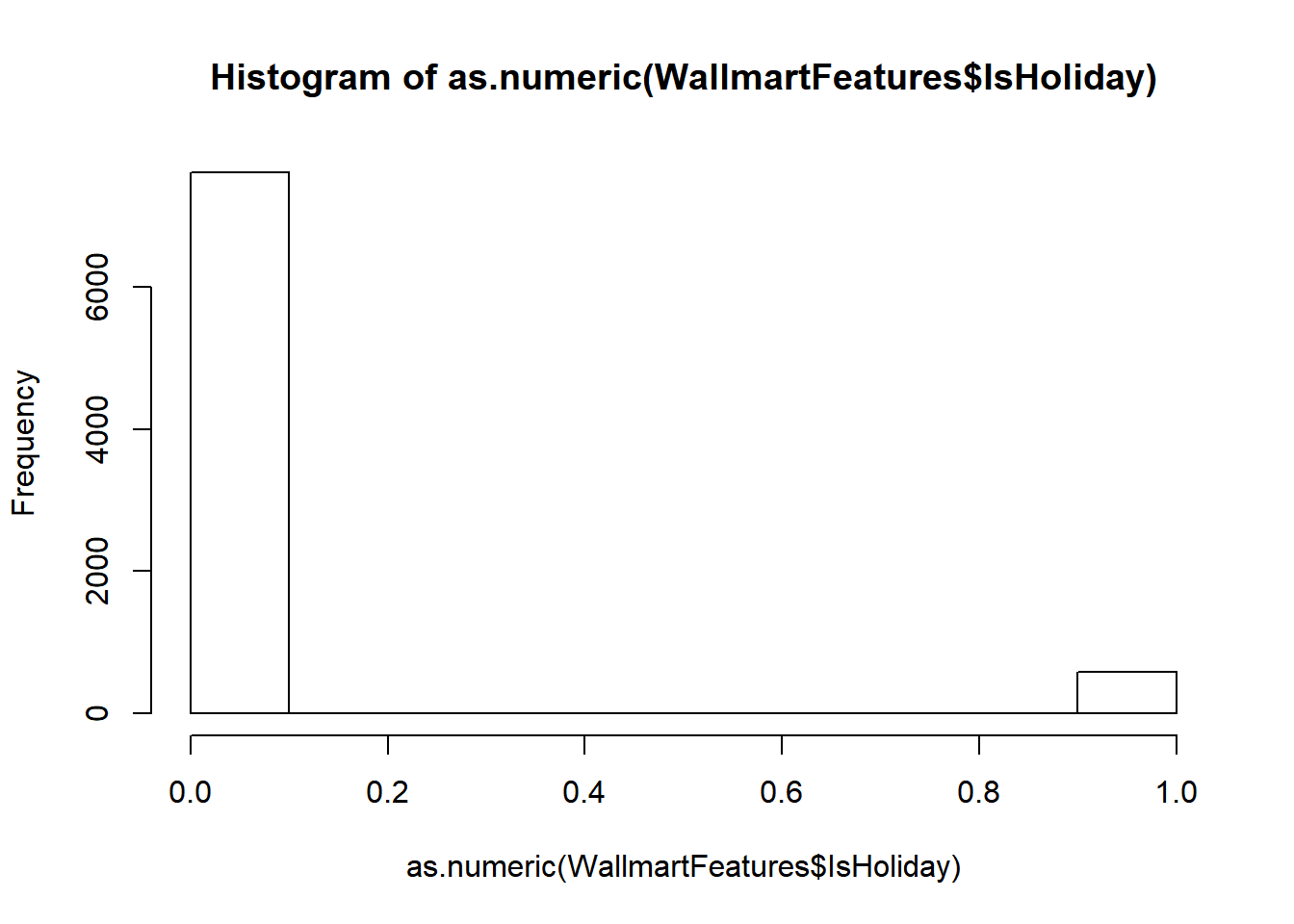


The CPI distribution is not normal and further investigation is needed to justify normalization. Without understanding the nature of inflation, no transformation can reasonably applied.

### Unemployment Rate Distribution

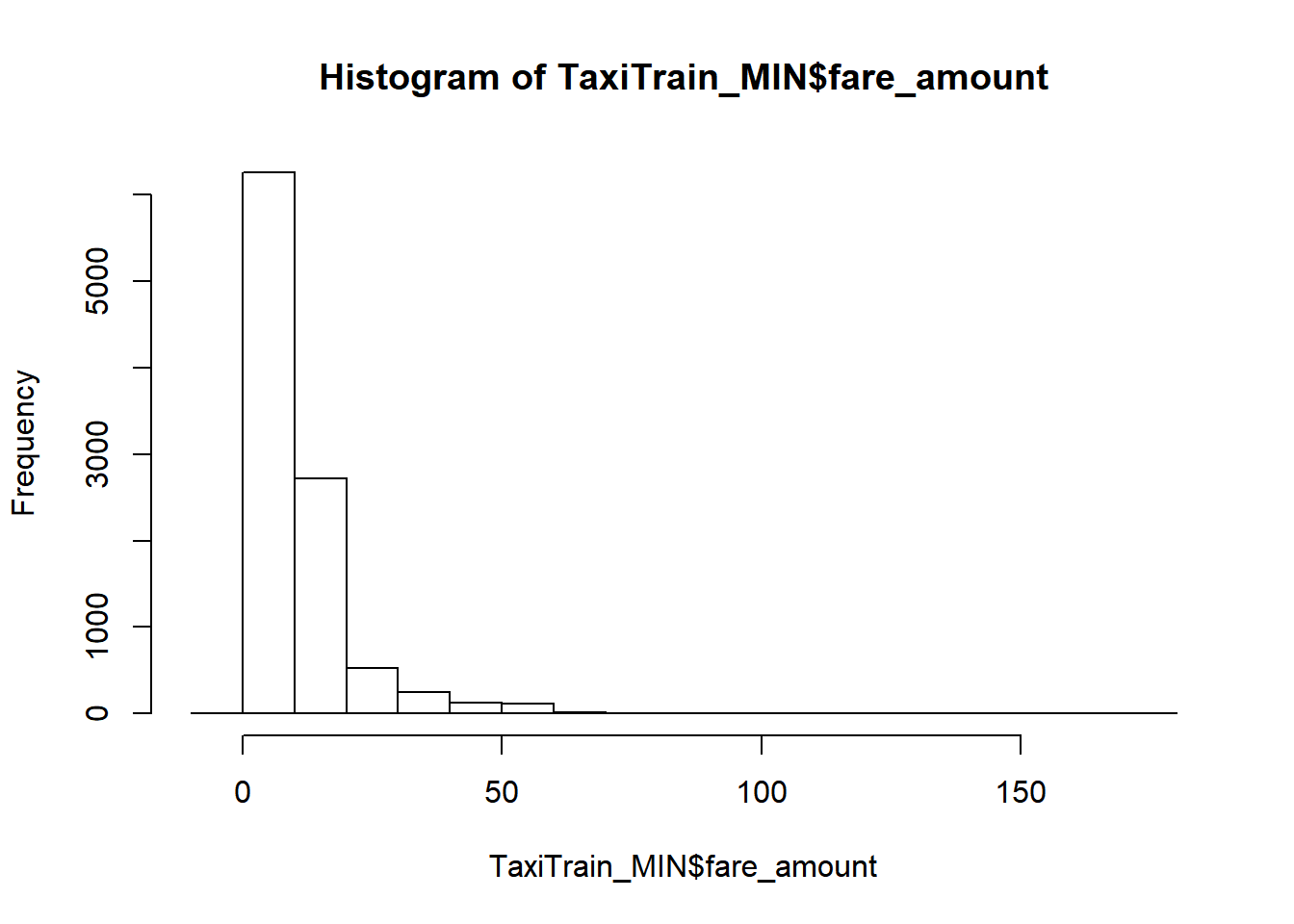


### Is Holiday Distribution

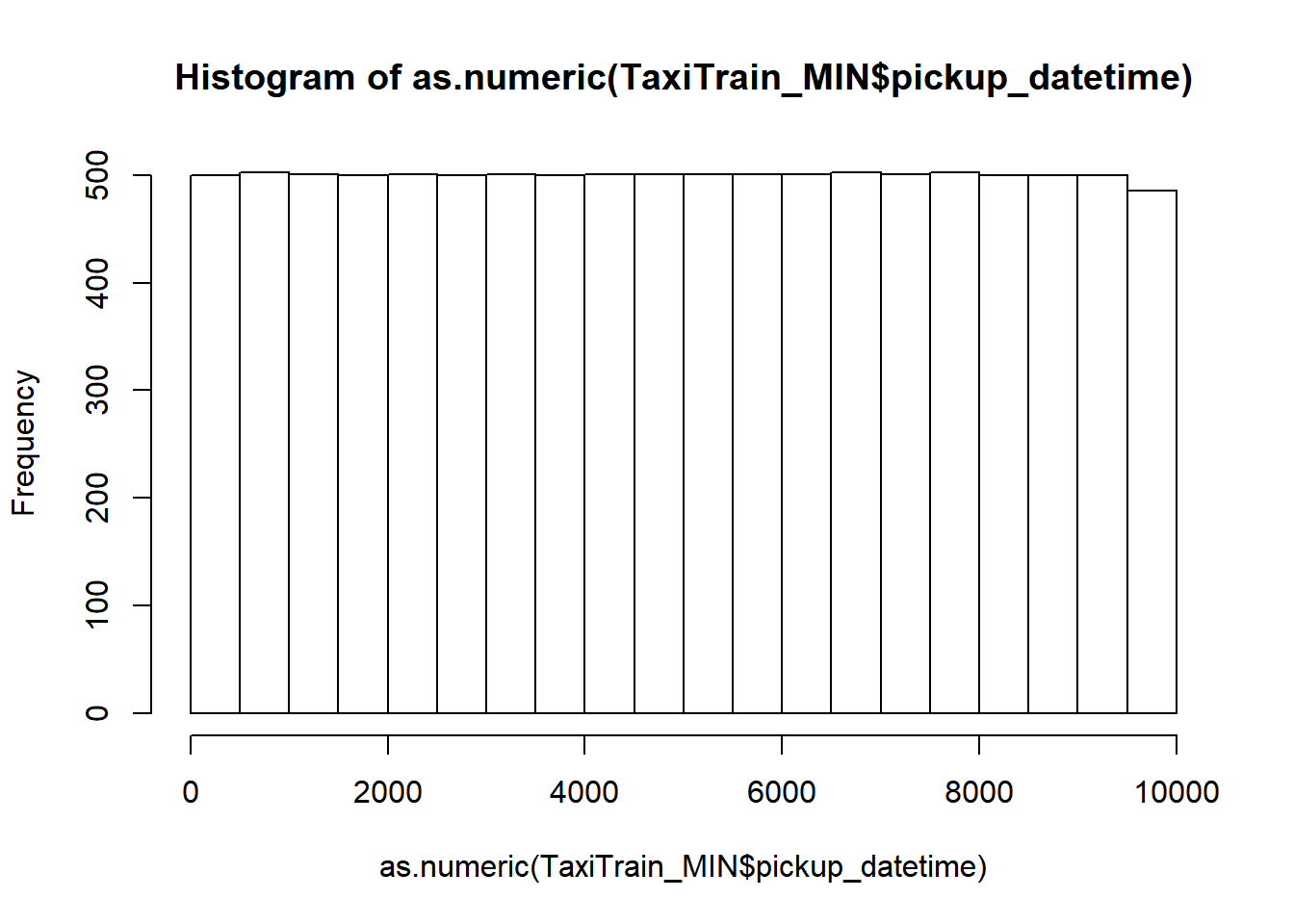


## Taxi Fare

### Fare Amount Distribution

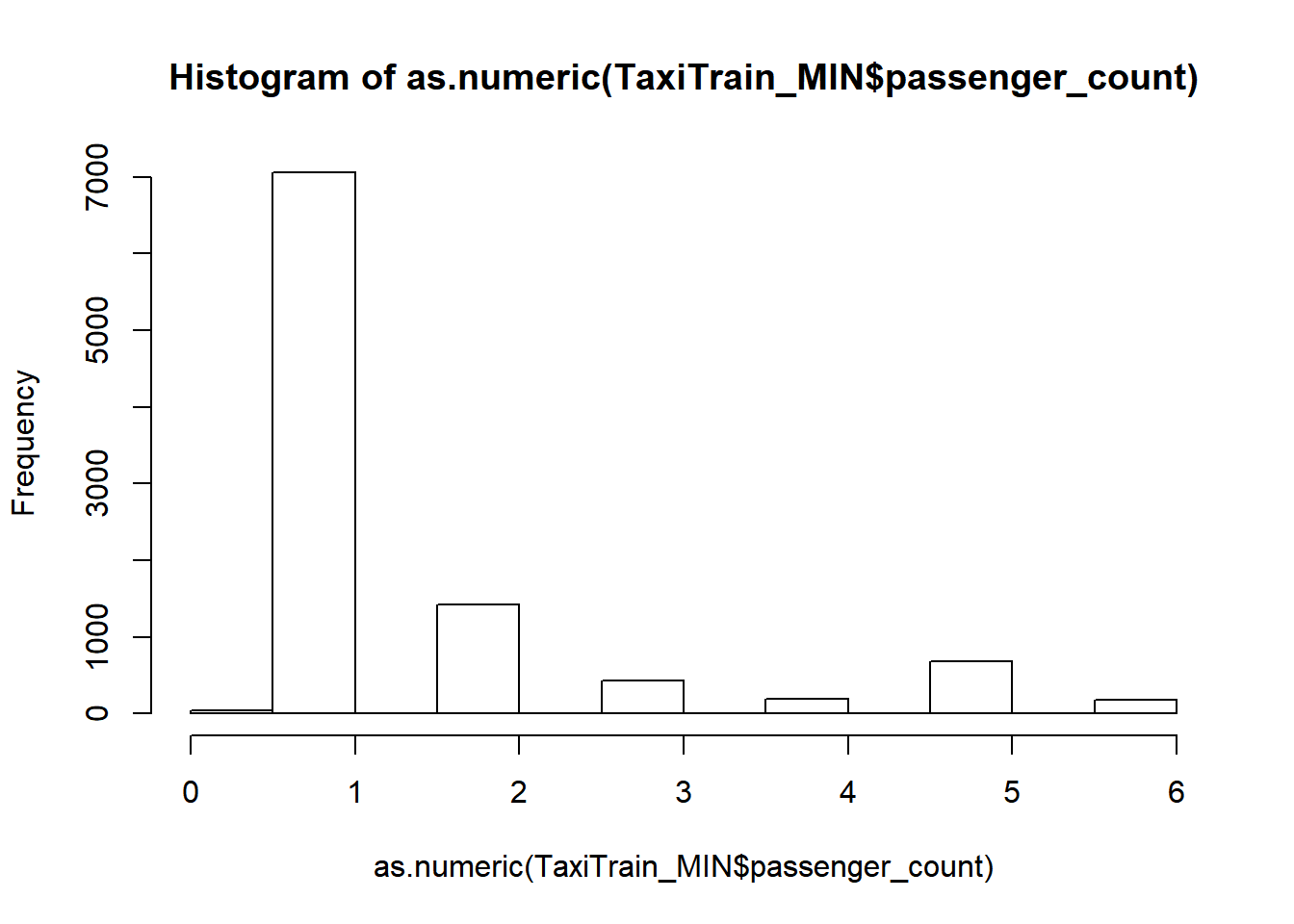


### Pickup Date Distribution



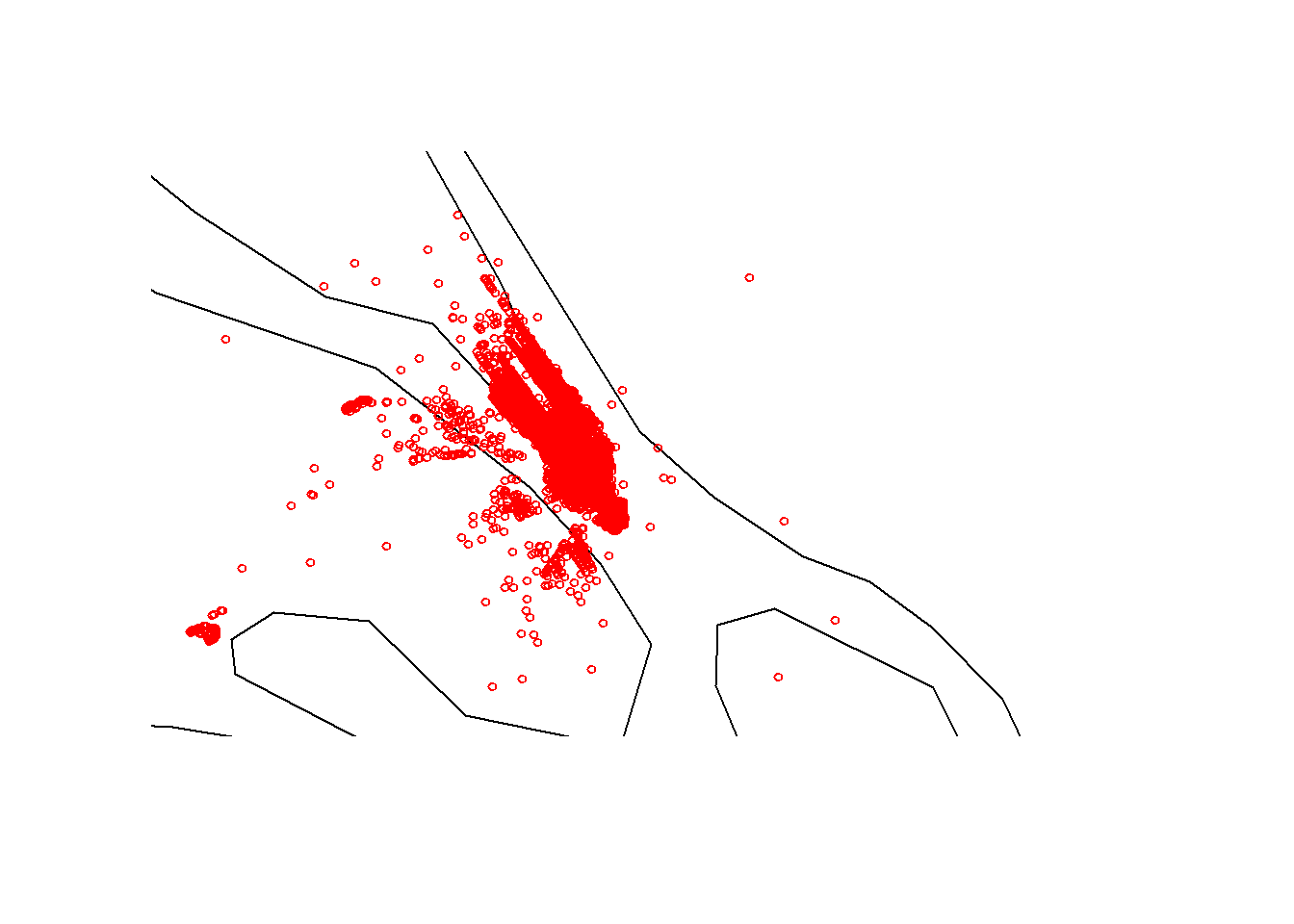
Equal data points exists for the range 2009/01/04 to 2015/06/11

### Passenger Count Distribution



### Pickup and Drop-off locations

Map of New York City



# Data preparation and feature engineering

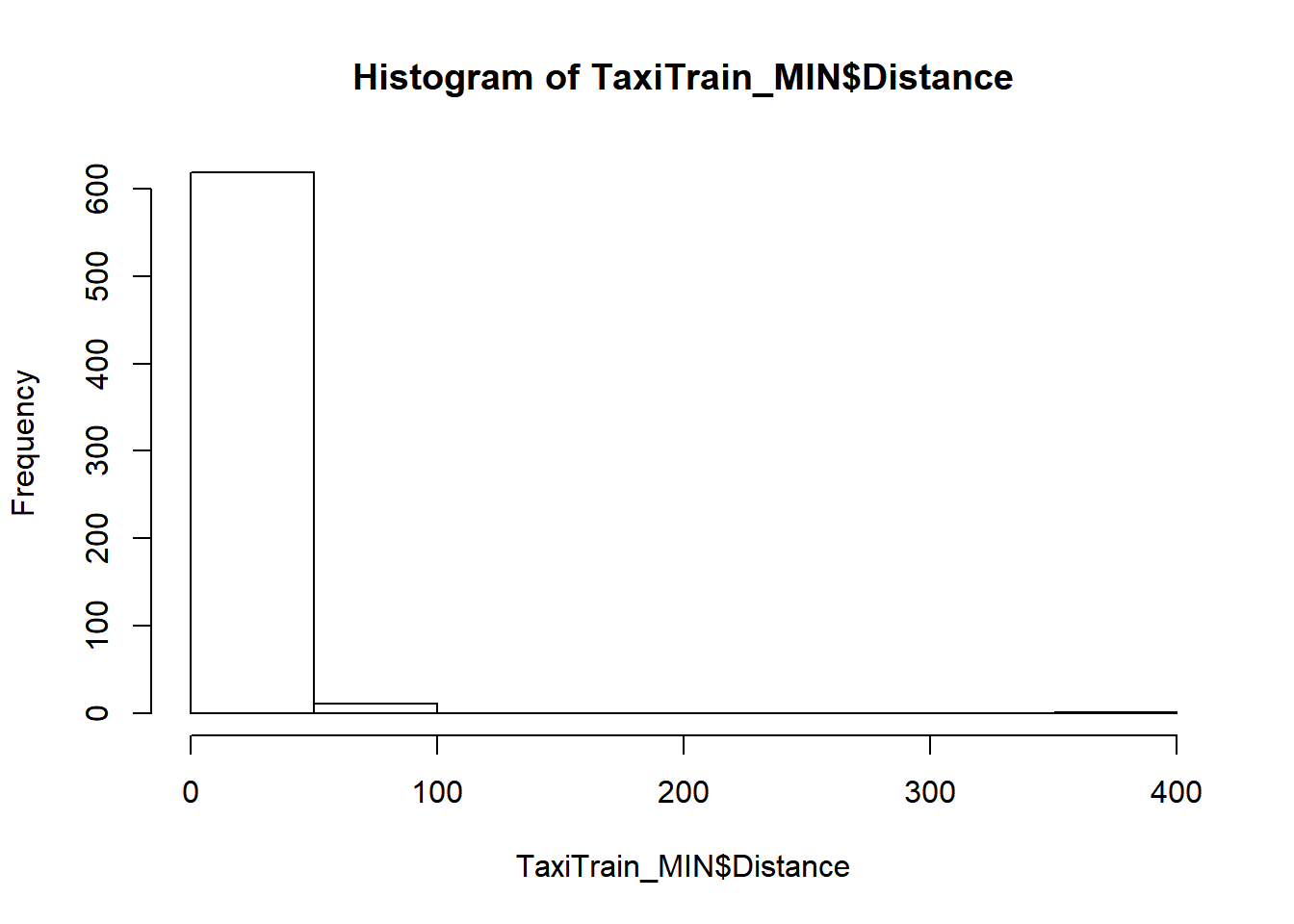
## Analysis of Missing Values

The Walmart dataset is aggregated by week number and hence has no missing values. As stated before, entire subsets of Taxi Fare data were excluded for this solution, also resulting in no missing values.

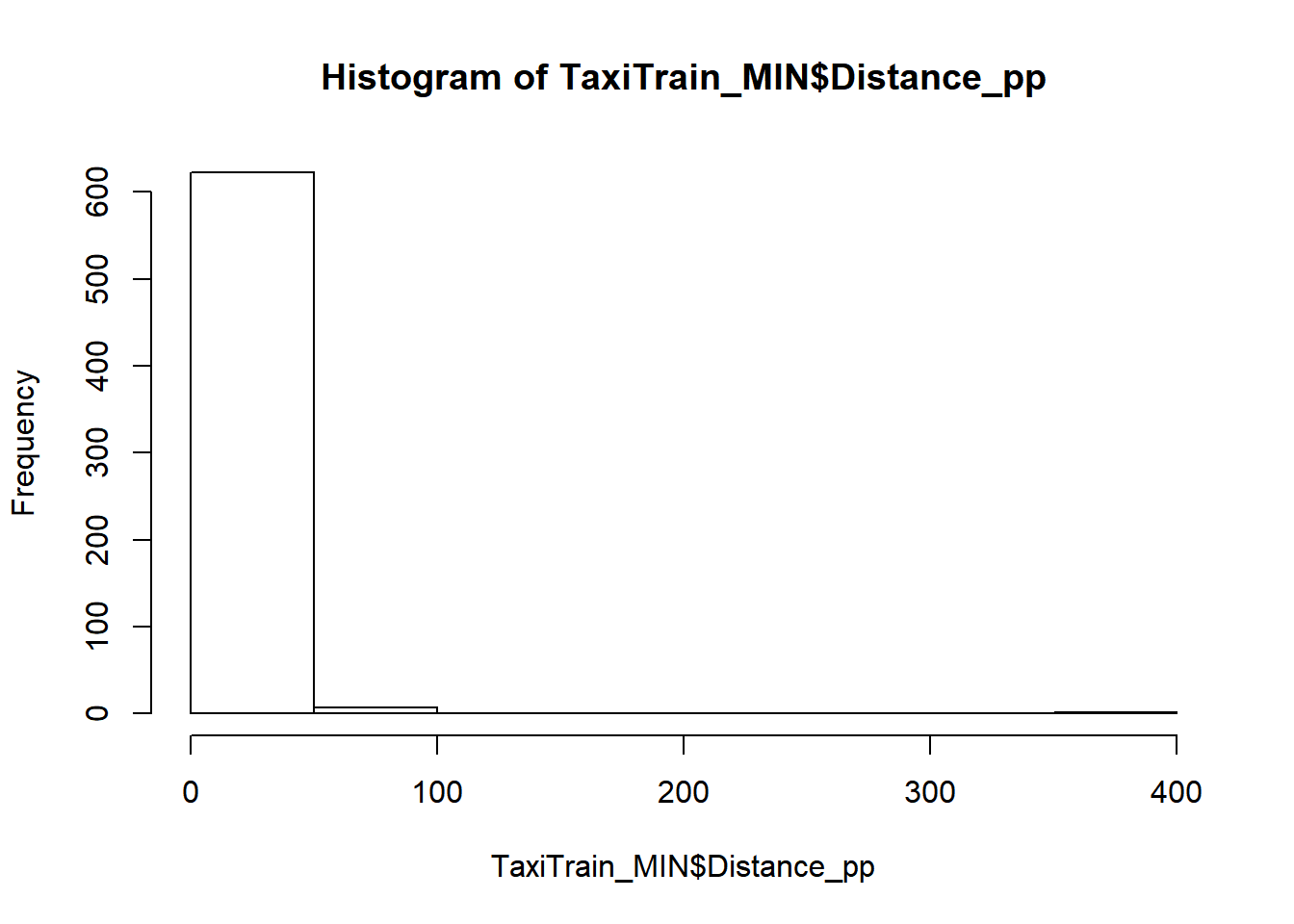
## Feature Engineering

The Walmart dataset is averaged by week number (1 to 53). Distance and fare per person was computed and added to enhance the Taxi Fare set.

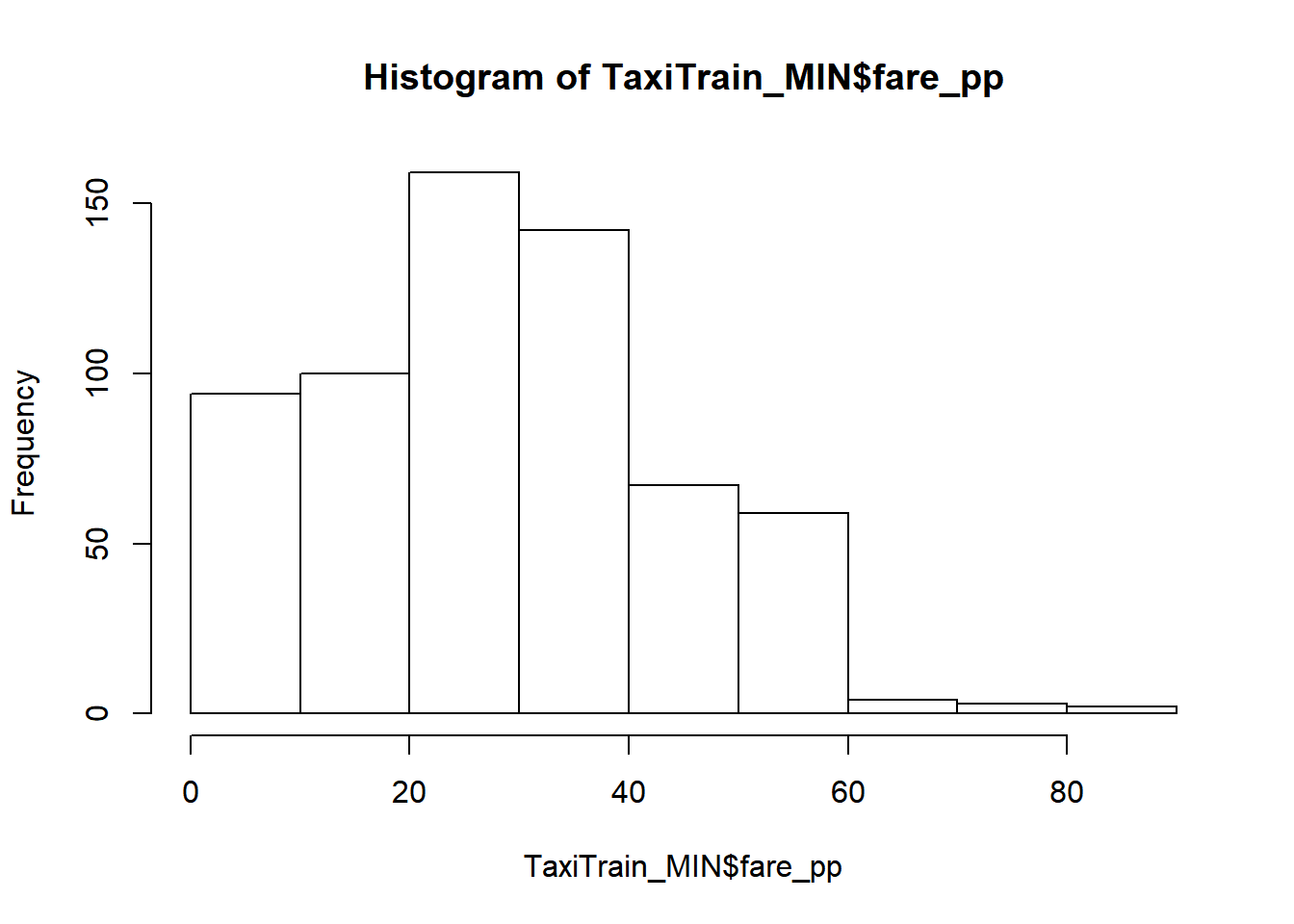
### Taxi Fare Distance distribution



### Distance per person distribution



### Fare per person distribution



# Problem Statement

Fluctuations in unemployment, CPI, temperature and fuel prices will cause changes in the number of passengers that share a taxi. Furthermore, unemployment, CPI, temperature and fuel prices are positively correlated to the number of passengers per ride.

# Methodology

Principal Component Analysis will be used to explore the relationship between all pairs of attributes and show their relationship graphically by plotting each resulting component’s rotation on a unit circle. The distance between attributes indicates their correlation: point that are closer together indicate a stronger positive correlation. Points that are further away from the origin describes bigger proportional variance in the data, hence to reduce noise in the data set, points in the centre of the unit circle could be excluded. A Scree Plot is a simple line segment plot that shows the fraction of total variance in the data as explained or represented by each component. The components are ordered, and by definition are therefore assigned a number label, by decreasing order of contribution to total variance.

Multiple decision trees will be used to predict the number of passengers based on different combinations of features. The tree with the lowest relative error will be chosen as the champion model.

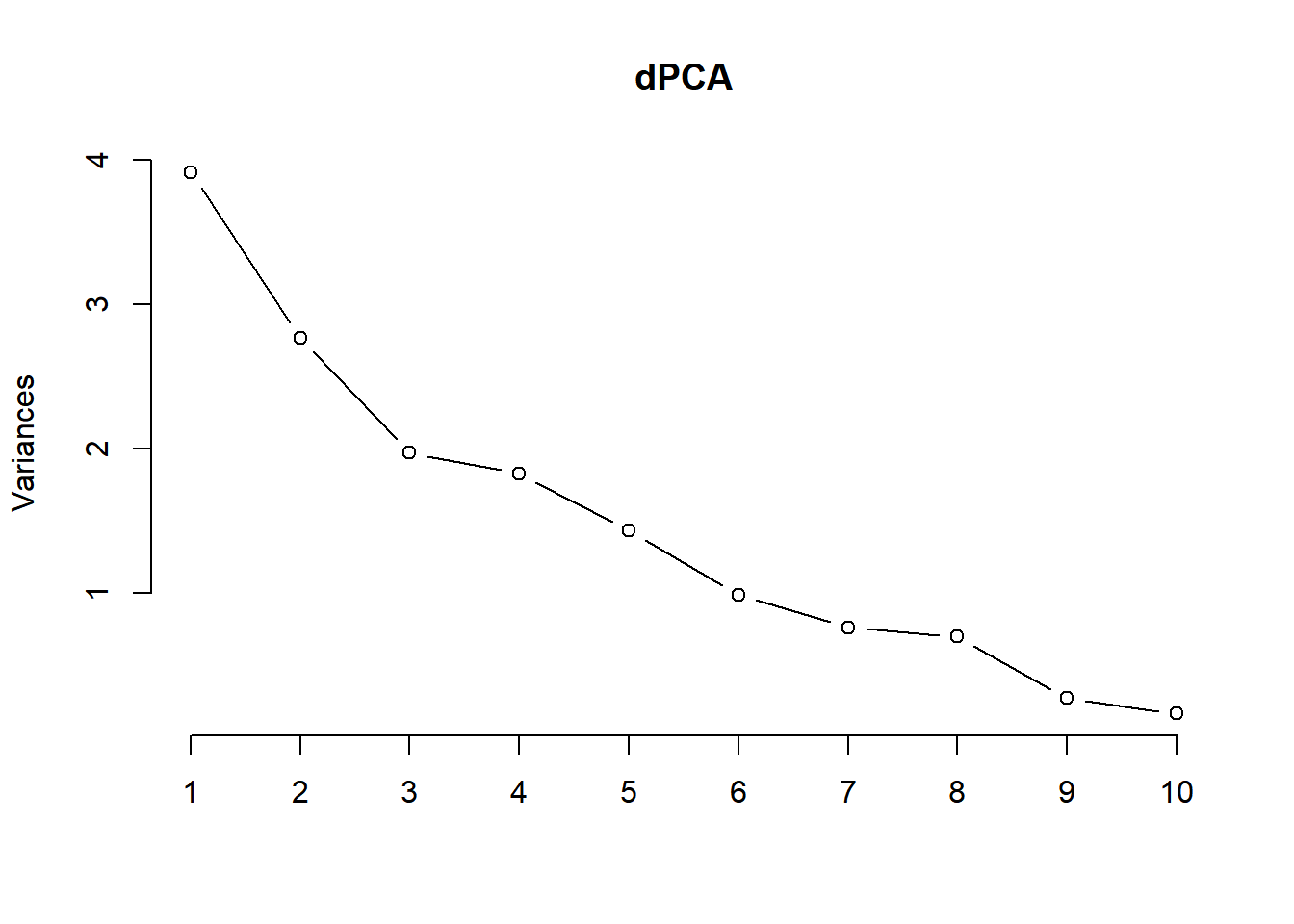
Ideally the component resulting from the PCA should be used in the model, instead of the original feature. A more accurate algorithm could be used as primary predictor alongside a secondary algorithm that facilitates interpretability. Unfortunately, given the time constraints, this method falls outside the scope of this solution.

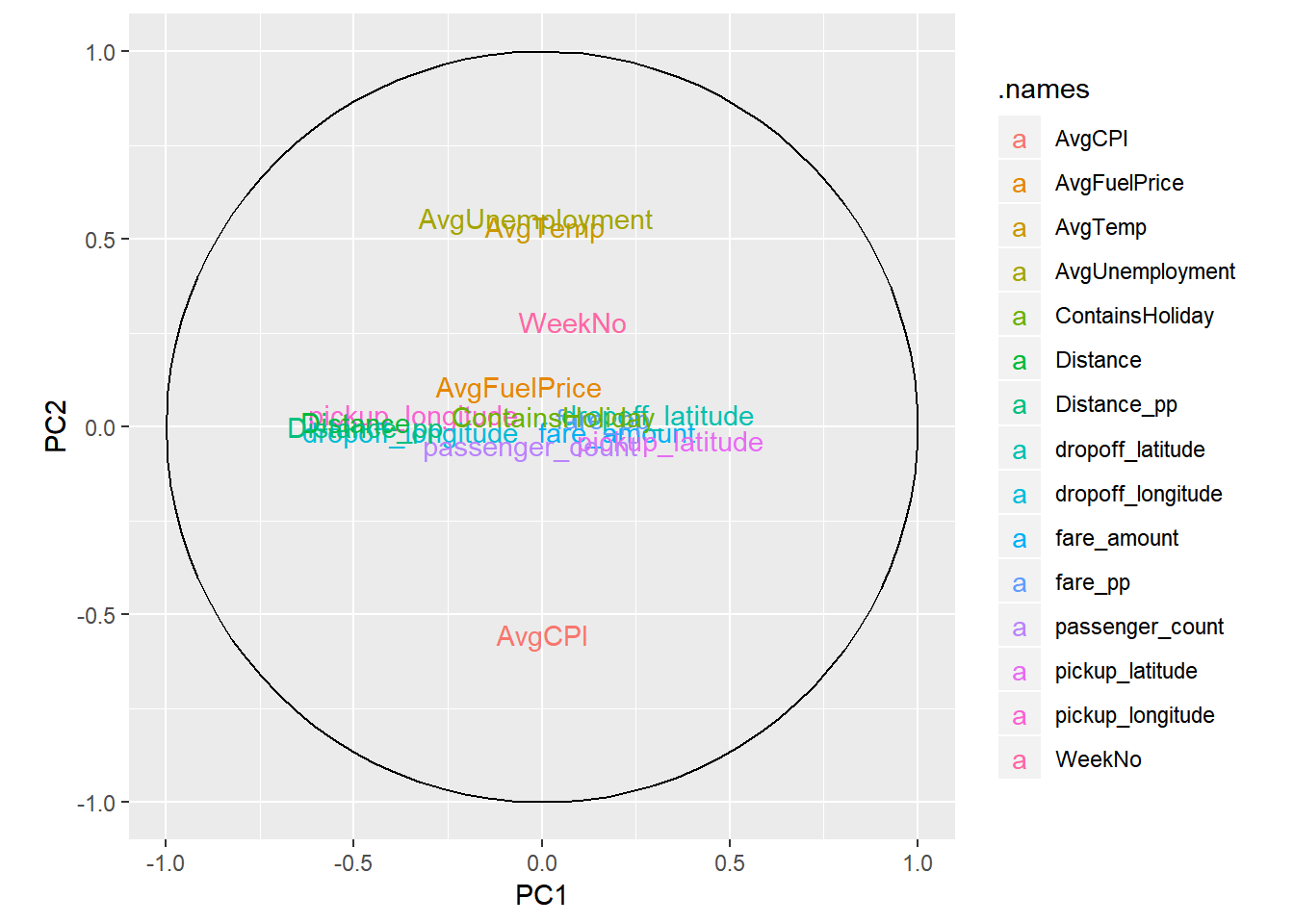
# Solution

## Feature Selection based on PCA

### First Iteration

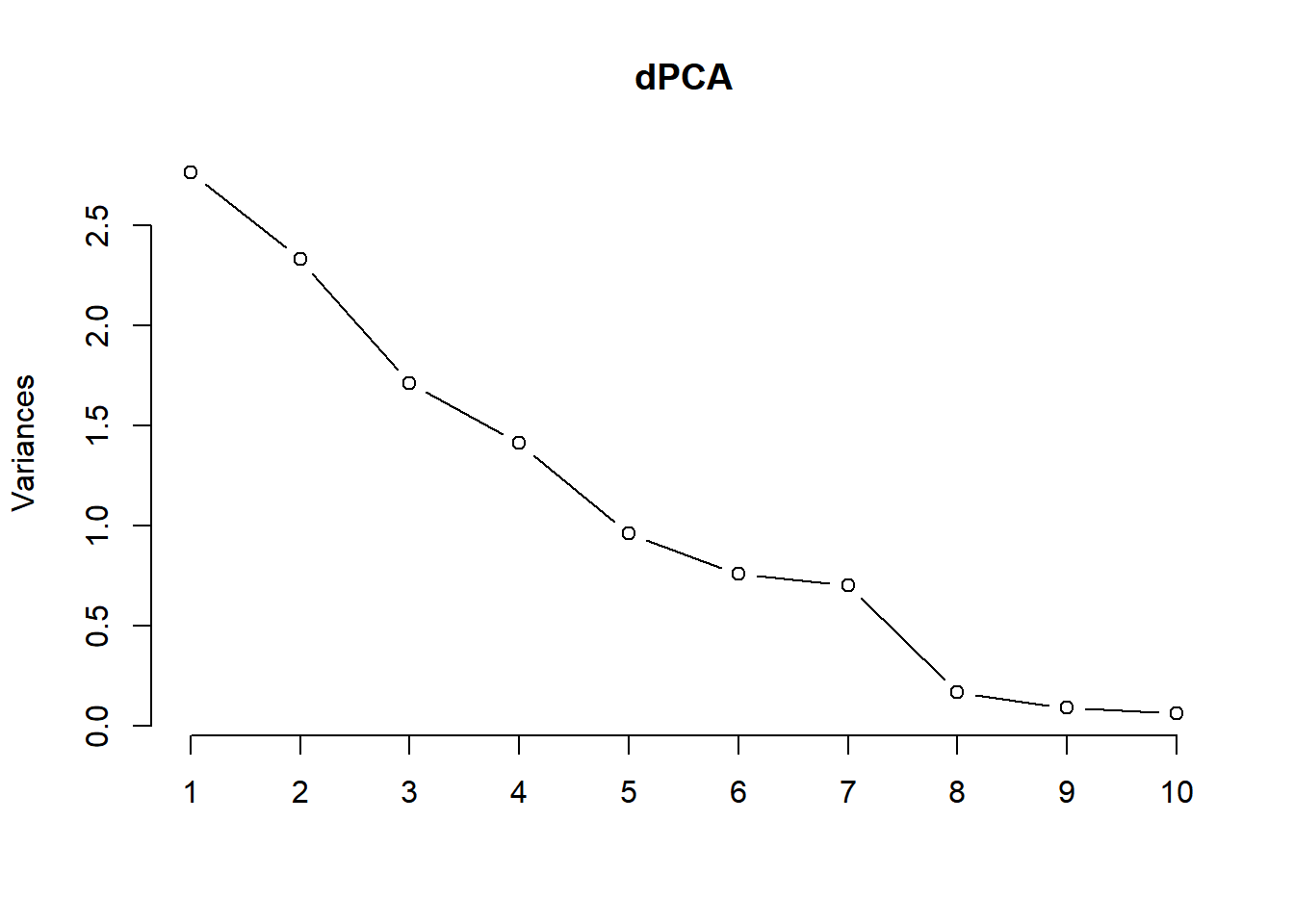
All attributes were included in the analysis.

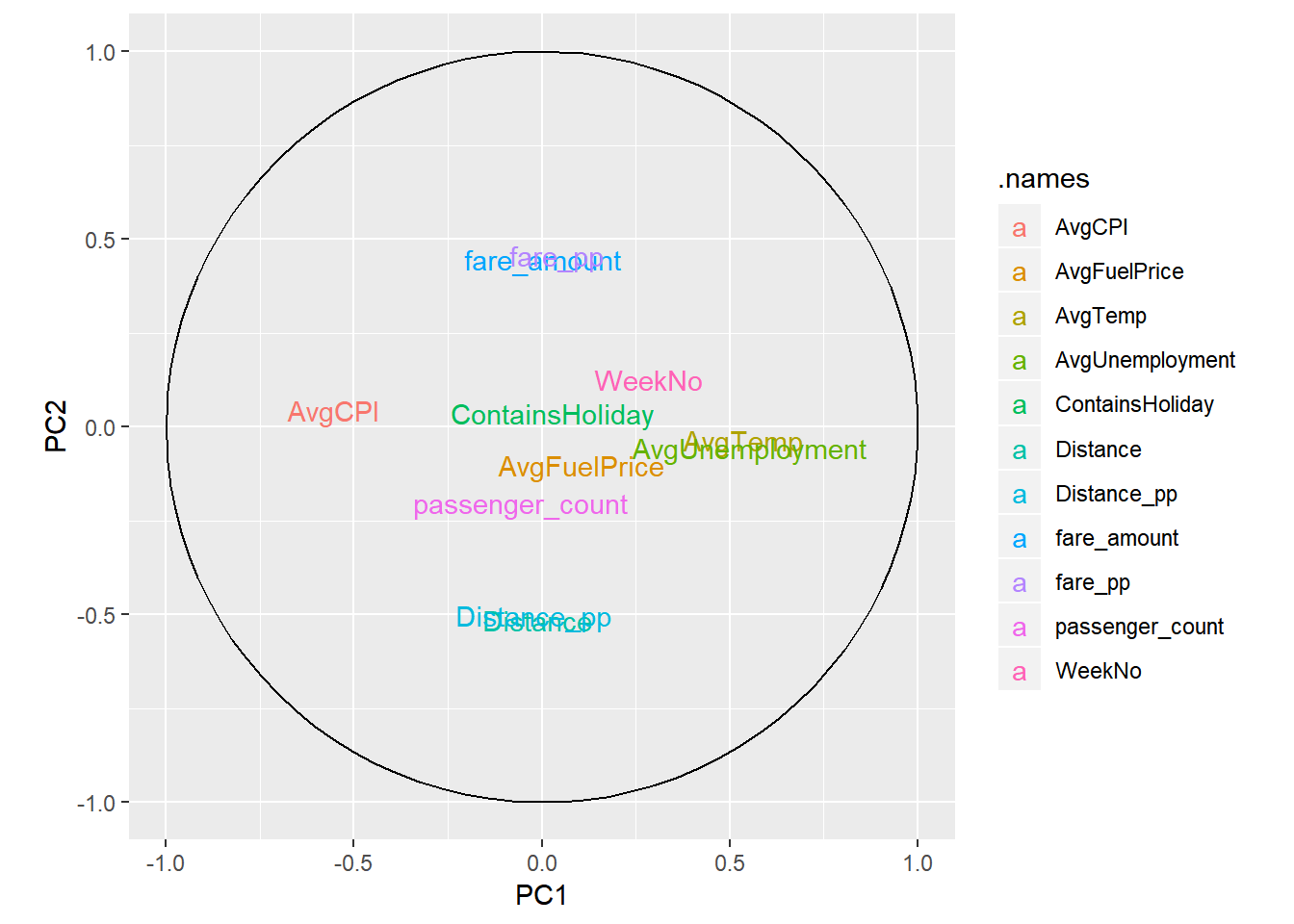




The analysis indicates that the pickup and drop off latitude and longitude are insignificant and can be excluded.

### Second Iteration





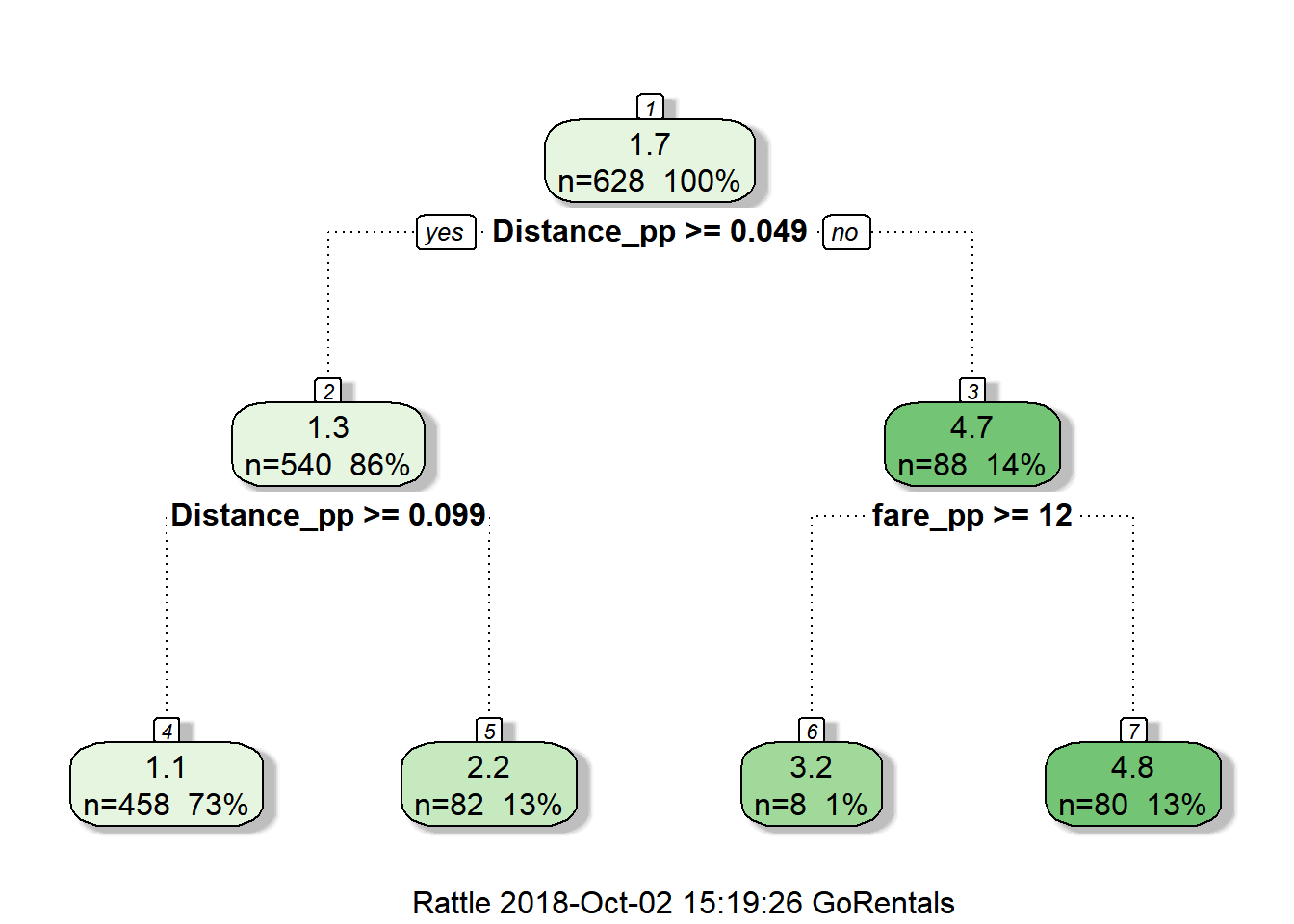
Fare amount per person and the total fare amount are describing the same variance in the data. Similarly, distance per person and total distance are also strongly correlated. Only one of each should be used and therefore Total Distance and Total Fare Amount is excluded.

The final set of attributes includes only the Week Number, Number of Passengers, Distance per person, Fare amount per person, Average Temperature, Average Fuel Price, Average CPI, Average Unemployment and the Boolean indicator Contains Holiday.

## Decision Trees

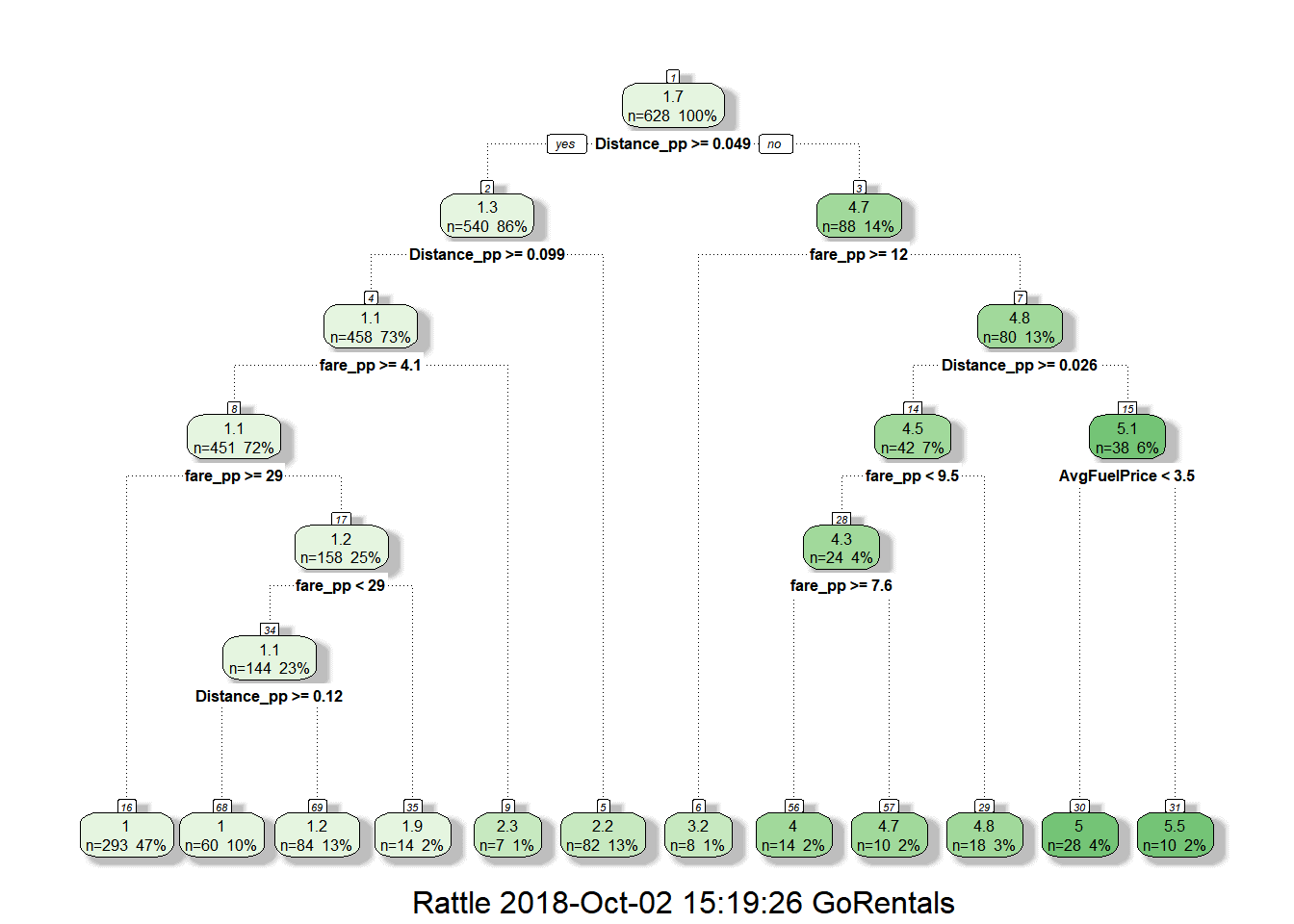
### First Iteration

No control parameter were specified and all attributes were included. The relative error is 0.1065536 for this model. Only Distance per person and Fare per person were important.



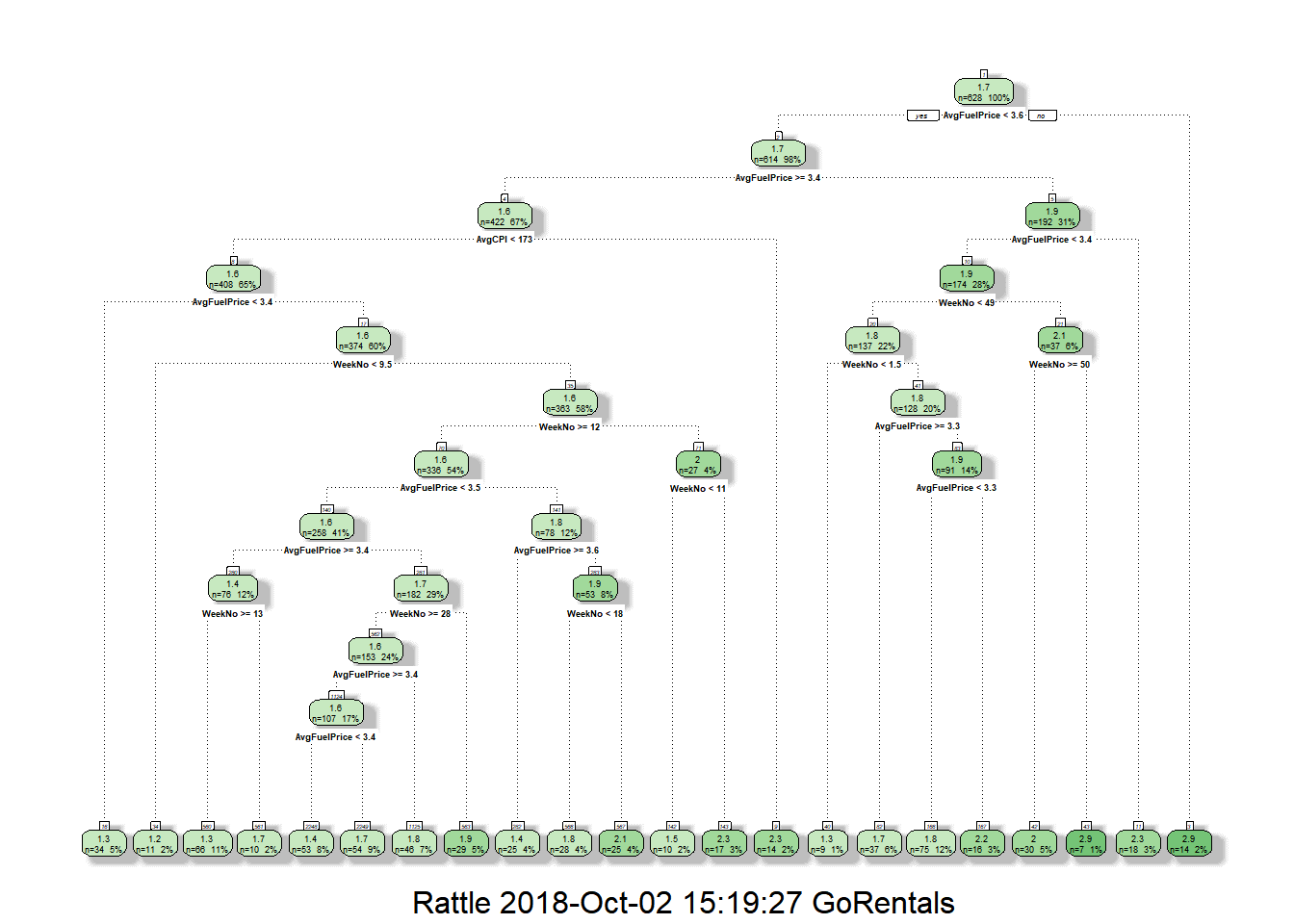
### Second Iteration

The control parameter was set at 0.001 which improved the relative error by decreasing it to 0.07116911. Again the only importance attributes were Distance per person and Fare per person.



### Third Iteration

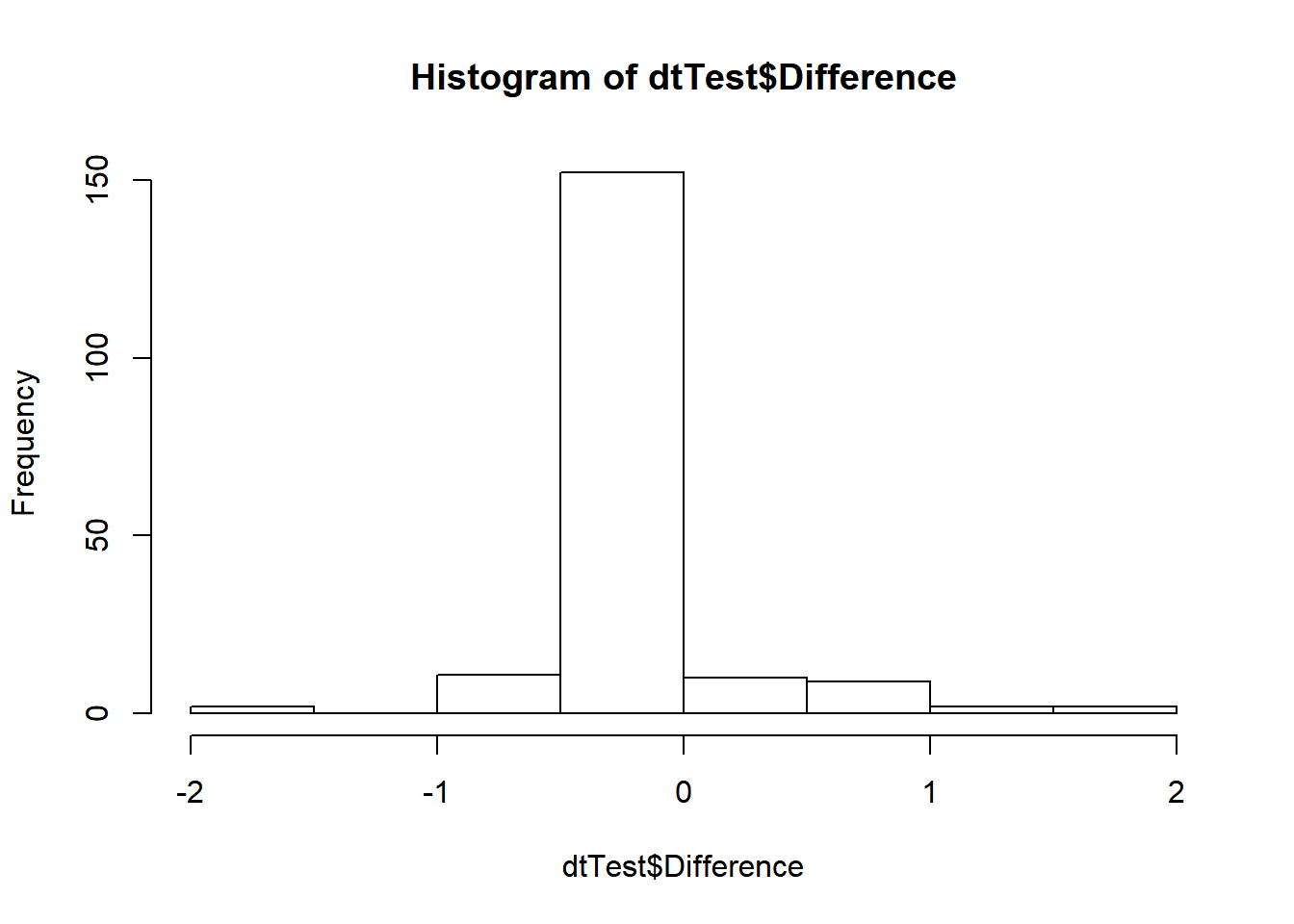
Since none of the environmental attributes were important enough to be included in the previous two models, Distance per person, Fare Amount per person and Average Temperature was excluded in an attempt to finally disprove the problem statement. This iteration has a relative error of 0.9215237, which is much higher than iteration one or two.



# Model Fitting and Analysis

Iteration Two’s decision tree had the lowest relative error and is therefore chosen as the champion model for this solution. This model can predict the Number of Passengers with an 86% accuracy. To further evaluate the model, the difference between the predicted an actual Number of Passengers were calculated for each test observation. These differences are normally distributed and therefore acceptable.

The hypothesis has been disproven: environmental attributes are not indicative of changes in the number of people that share a taxi. Given the limitations and assumptions of this solution, Fuel price, CPI, Unemployment and Temperature has no relevant correlation and does not cause fluctuations in commuters’’ behaviour. However, within the boundaries of this solution’s assumptions, no definitive conclusion can be inferred.



# Suggestion for improvements

The Walmart dataset is not good enough to describe the environmental attributes for New York City. A more accurate dataset should be researched that’s specific to New York City and falls within the same date range as the Taxi Fare dataset.

More algorithms should be compared to find the champion model, one decision tree is too simplistic to infer any conclusions from the current data.

None of the visualisation were designed properly, extra time should be spent on choosing the best type of graph and setting the appropriate axis, colours, and data labels.

Code Repository

<https://github.com/LouiseDeBeer/BRICS2018>

# Cognitive Solution

A separate report will be submitted to describe our team’s Day Two challenge. The solution recommends movie directors that a user would like based on an analysis of their social media posts. The solution runs on the IBM Cloud and uses Watson’s Personality Insights API to output Consumption Preferences from a given Twitter account. The Consumption Preferences are then used with a Watson Machine Learning Model to recommend movie directors. The Watson Machine Learning Model uses an augmented IMDB dataset that includes movie metadata such as genres, keywords and Facebook likes.

# About our team



Nabeel Valley matriculated in 2013 from Al Ghazali College in Pretoria. He went on to complete a Bachelor of Engineering degree from the University of Pretoria in 2017, where he was also the chairperson for the Muslim Students Association. Nabeel has experience as a university tutor, an engineering assistant and recently moved over to software development. He’s currently working as a Developer Advocate at IBM. Nabeel is a hobbyist photographer and graphic artist with a love for learning new technologies.

[Nabeel on LinkedIn](https://www.linkedin.com/in/nabeelvalley)



Louise de Beer has over a decade experience as a full stack .Net developer and has been focusing on data science and analytics since 2016. Louise completed a BSc. Applied Mathematics degree in 2008 from the University of Johannesburg. She’s an amateur violinist and mom of two. Louise is currently working as a data scientist at Investec.

[Louise on LinkedIn](https://www.linkedin.com/in/louise-de-beer-a9630115)