# Regression Project

Jarius Hamid

2024-08-26

```
knitr::opts_chunk$set(echo = TRUE)
```

## Loading all of The Libraries and Data

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats 1.0.0     v stringr 1.5.1
## v purrr   1.0.4     v tibble  3.2.1
## v readr   2.1.5     v tidyr   1.3.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(e1071)
library(xtable)
library(psych)
```

```
##
## Attaching package: 'psych'
##
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

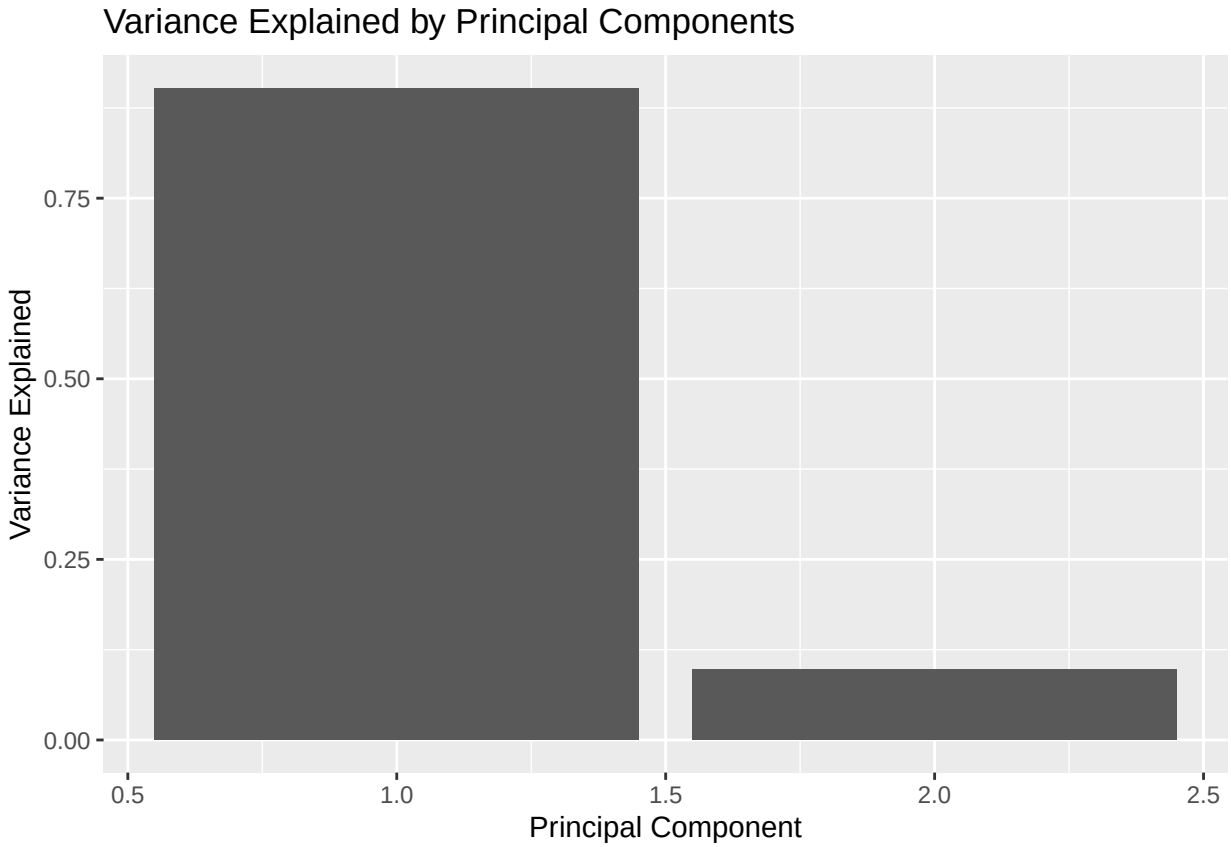
```
library(fixest)
library(stringr)
library(forcats)

transit_data <- read.csv("r_reg.csv")
rain_and_temp <- read.csv("rain_and_temp_data.csv")
miles_and_stops <- read.csv("route_milesandstops.csv")
```

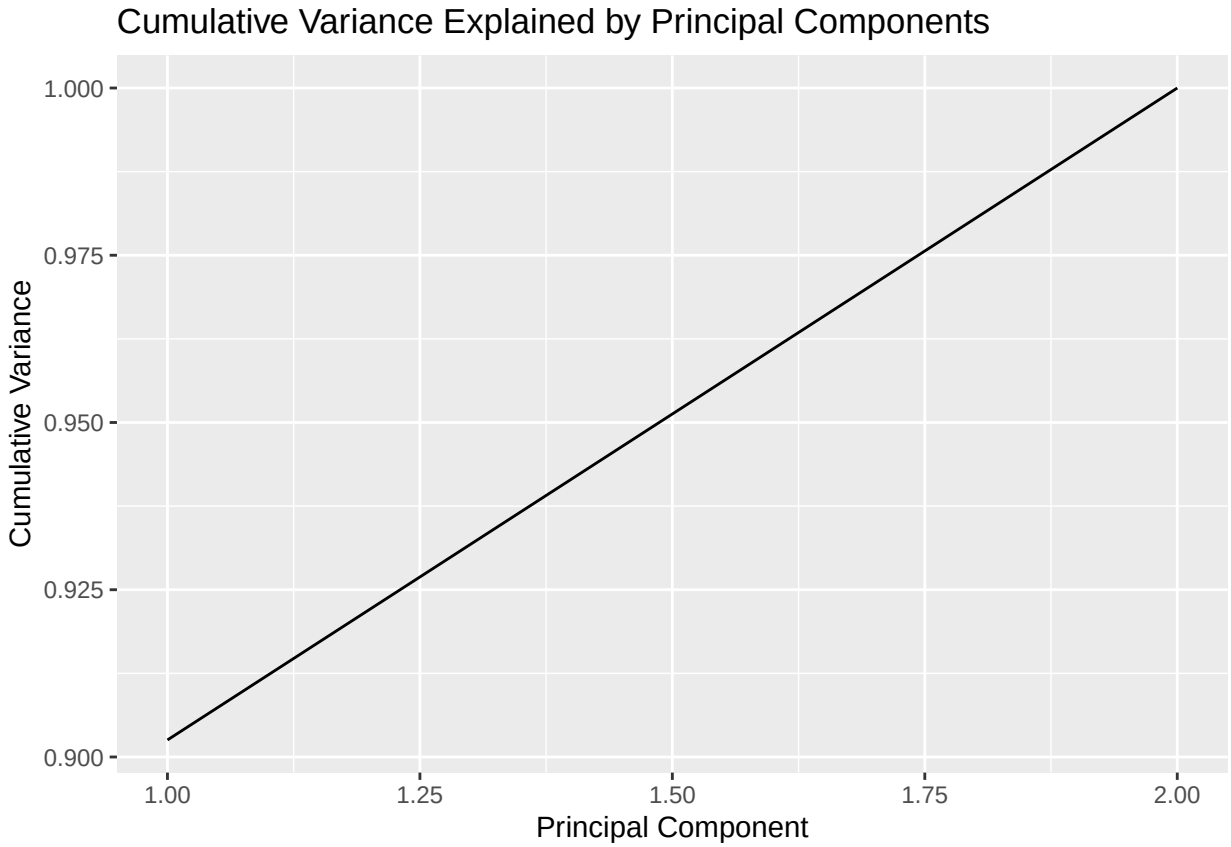
# Data Manipulation

# Principle Components Analysis

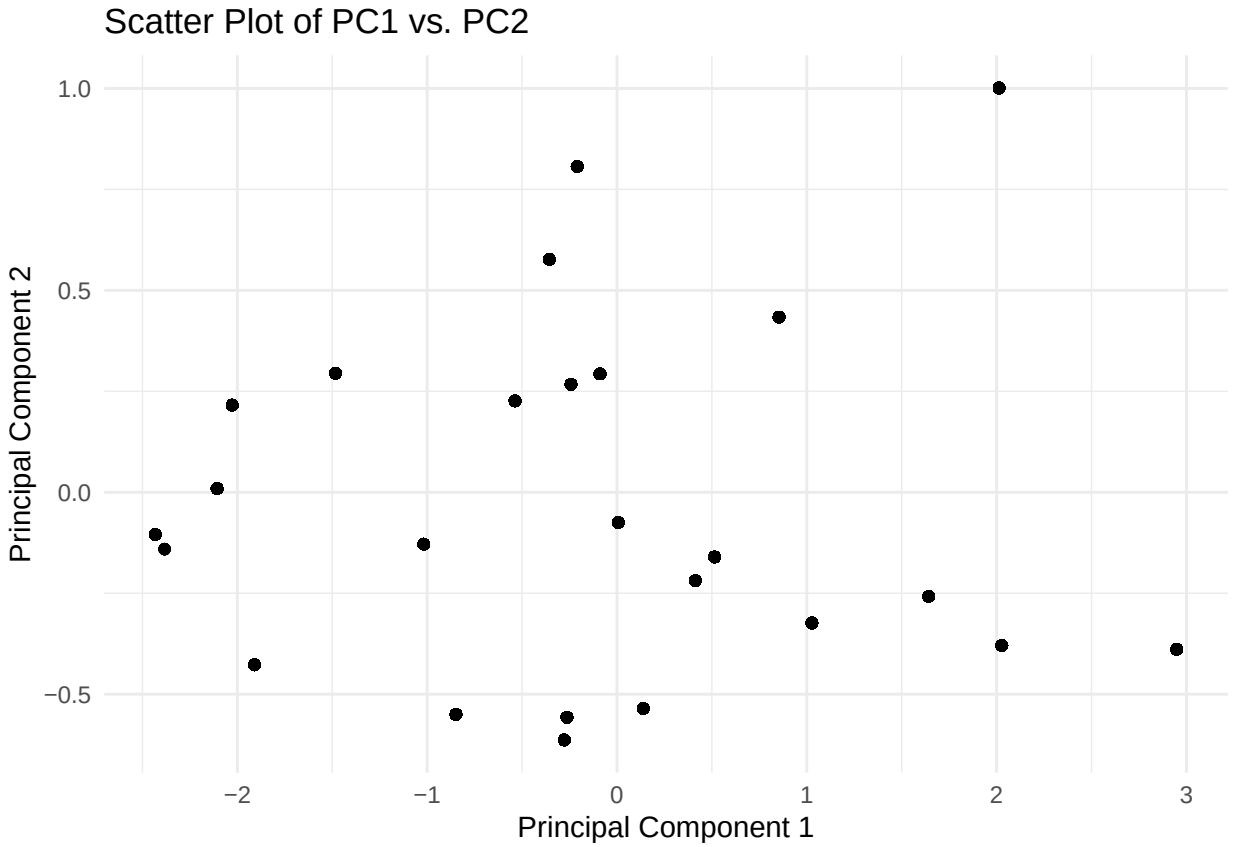
```
## Importance of components:
##                           PC1     PC2
## Standard deviation     1.3435 0.44149
## Proportion of Variance 0.9025 0.09746
## Cumulative Proportion  0.9025 1.00000
```




```
##                       PC1        PC2
## miles_roundtrip 0.7071068  0.7071068
## stops           0.7071068 -0.7071068
```




```
##          PC1        PC2
## 1 -0.2774899 -0.6134051
## 2  1.0267975 -0.3236498
## 3  2.0133898  1.0011199
## 4 -0.3557549  0.5764949
## 5  2.0278193 -0.3794322
```



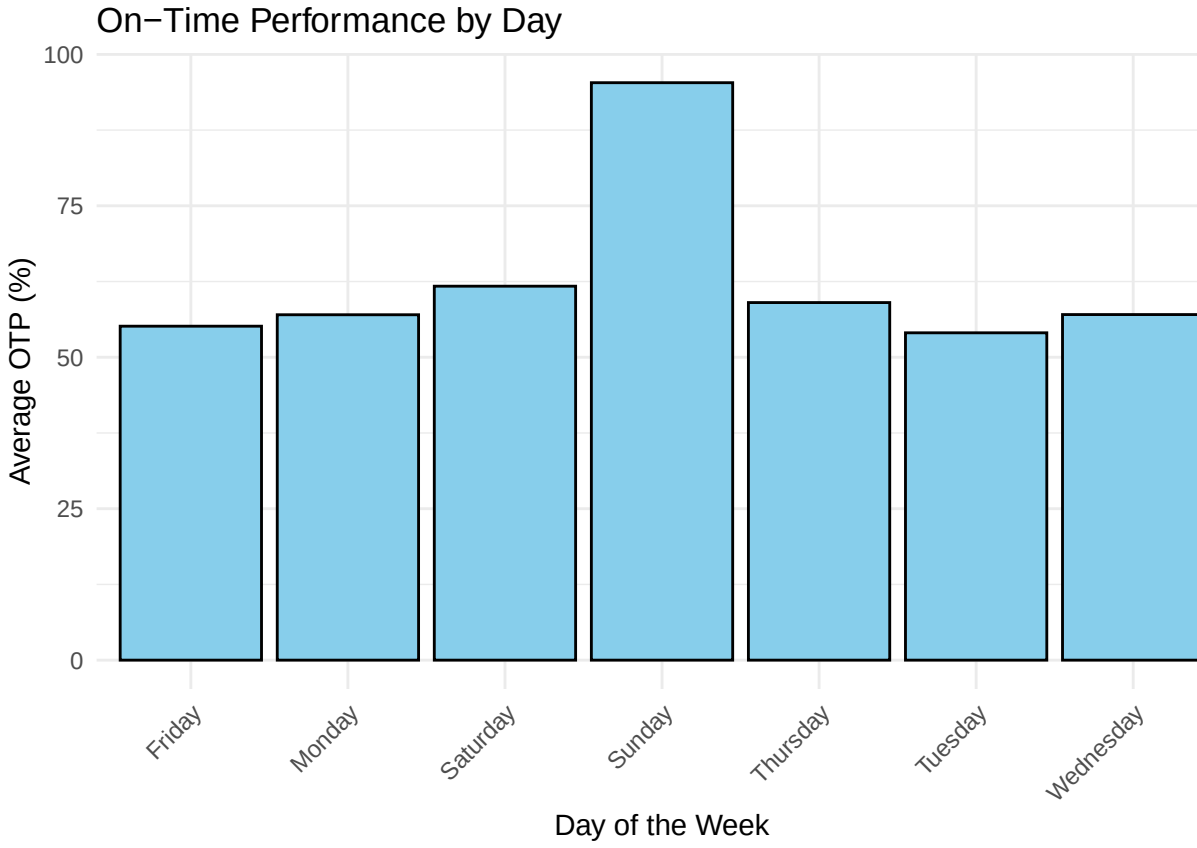

# Summary Statistics with Latex Ouput

```
## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf
```

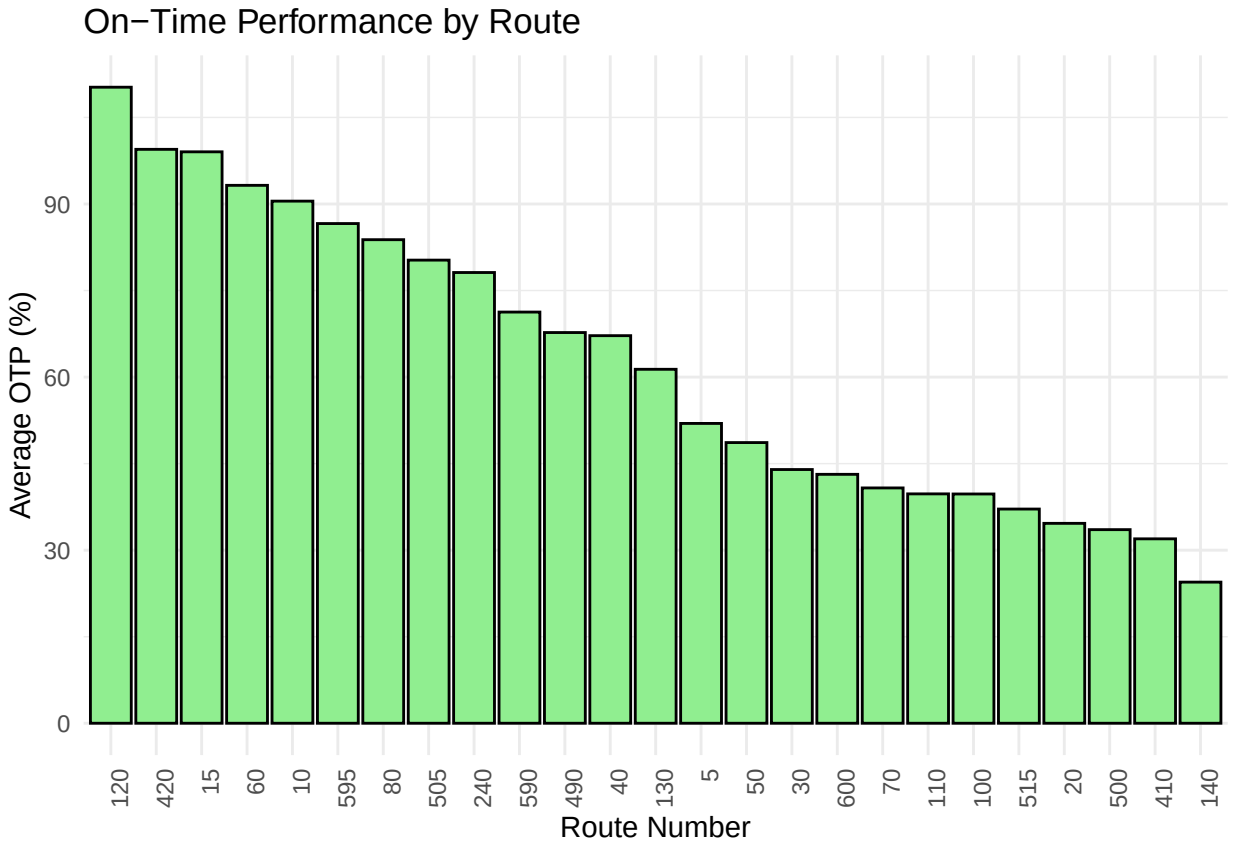
```
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf
```

```
## % latex table generated in R 4.4.2 by xtable 1.8-4 package
## % Wed Feb 26 21:19:17 2025
## \begin{table}[ht]
## \centering
## \begin{tabular}{lrrrrrr}
##   \toprule
## Variable & Mean & SD & Median & Range & Skew & Kurtosis \\
##   \midrule
## date &  &  &  & -Inf &  &  \\
##   day* & 4.02 & 2.07 & 4.00 & 6.00 & -0.02 & -1.36 \\
##   route & 218.80 & 219.87 & 110.00 & 595.00 & 0.76 & -1.16 \\
##   sum\_processed & 135.05 & 69.72 & 119.00 & 377.00 & 1.45 & 2.60 \\
##   on\_time & 63.32 & 15.27 & 64.59 & 100.00 & -0.56 & 0.37 \\
##   late & 29.37 & 16.95 & 27.37 & 100.00 & 0.74 & 0.53 \\
##   early & 7.31 & 6.82 & 6.10 & 57.14 & 1.53 & 4.30 \\
##   avg\_rain\_inch & 0.13 & 0.36 & 0.00 & 3.66 & 5.57 & 40.71 \\
##   upt & 227.04 & 341.16 & 136.00 & 13175.00 & 8.46 & 220.33 \\
##   avg\_sched\_dev & 3.74 & 2.98 & 3.20 & 48.17 & 3.12 & 19.00 \\
##   index & 1.00 & 0.64 & 1.00 & 12.65 & 3.43 & 34.04 \\
##   accident\_count & 83.72 & 21.18 & 84.00 & 121.00 & -0.23 & -0.51 \\
##   tmax & 84.56 & 7.36 & 85.00 & 49.00 & -0.91 & 1.16 \\
##   miles\_roundtrip & 22.85 & 10.11 & 23.03 & 39.70 & 0.23 & -0.41 \\
##   stops & 78.05 & 33.45 & 80.00 & 134.00 & 0.35 & -0.30 \\
```

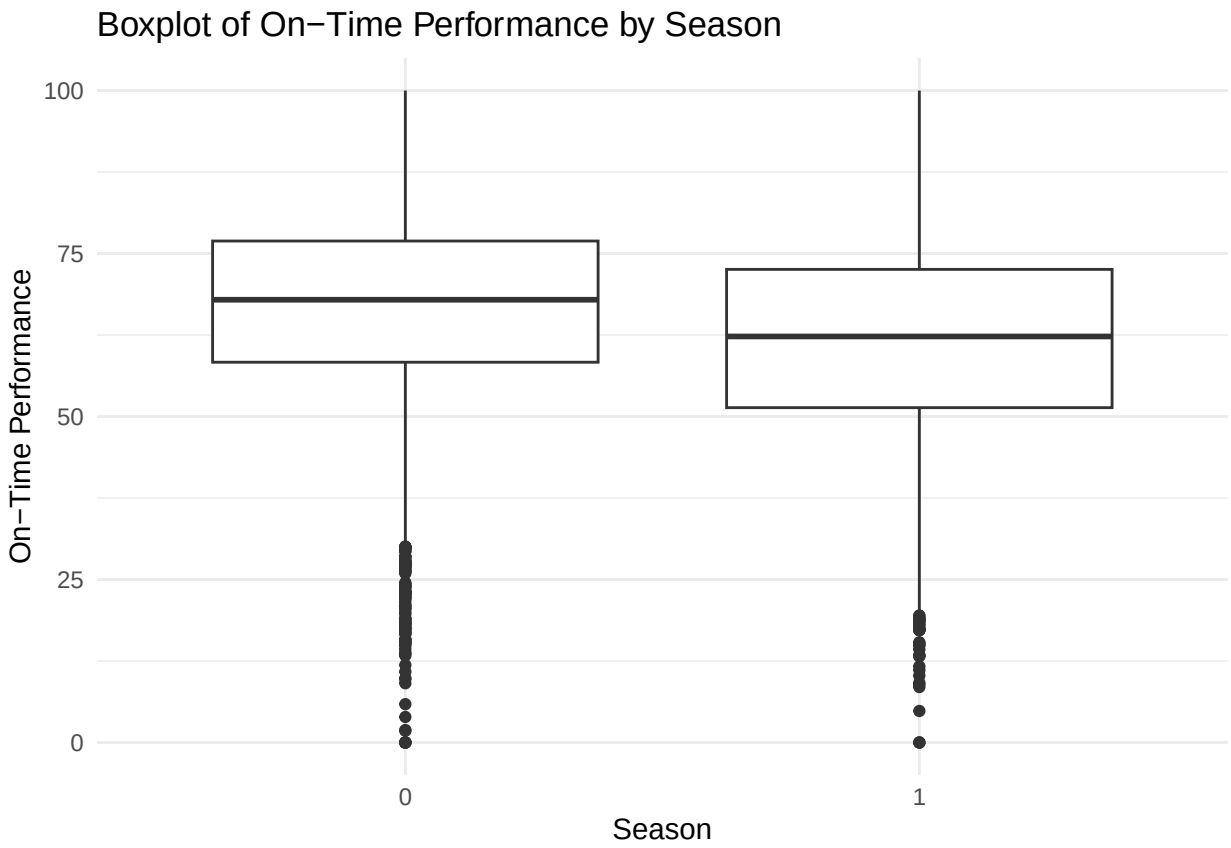
```
##   stops\_ftapart & 1633.79 & 483.32 & 1452.00 & 1962.00 & 0.55 & -0.46 \\
##   month & 6.15 & 3.69 & 5.00 & 11.00 & 0.23 & -1.32 \\
##   season & 0.61 & 0.49 & 1.00 & 1.00 & -0.45 & -1.80 \\
##   PC1 & -0.00 & 1.34 & -0.21 & 5.38 & 0.21 & -0.43 \\
##    \bottomrule
## \end{tabular}
## \caption{Summary Statistics}
## \label{tab:summary_statistics}
## \end{table}
```



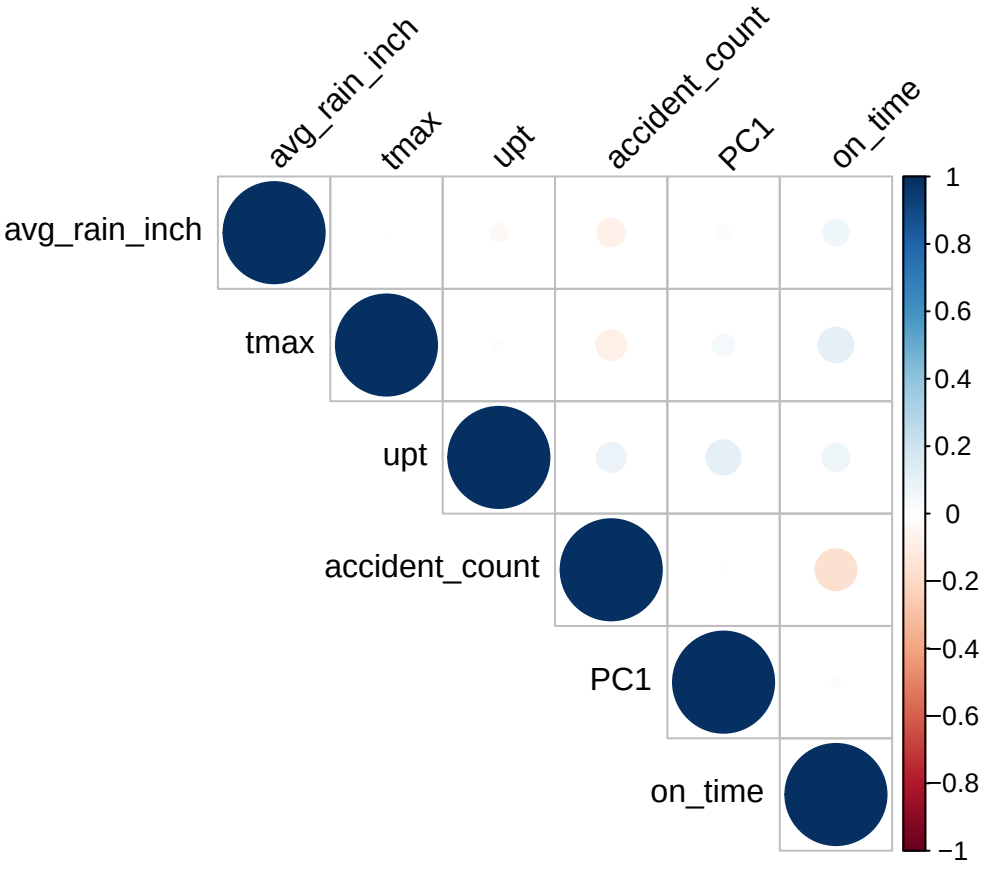

OTP BAR CHARTS


On−Time Performance by Route
Average OTP (%)
0
30
60
90
120
420
15
60
10
595
80
505
240
590
490
40
130
5
50
30
600
70
110
100
515
20
500
410
140
Route Number

## Scatter Plot Materix TESTING




```
## corrplot 0.95 loaded
```




## Explanatory Regression

```
## OLS estimation, Dep. Var.: on_time
## Observations: 9,848
## Fixed-effects: route: 25
## Standard-errors: Clustered (route)
##                      Estimate Std. Error   t value   Pr(>|t|)
## I(avg_rain_inch^2)  0.308413   0.118862  2.594720 1.5893e-02 *
## tmax                0.080768   0.053838  1.500219 1.4660e-01
## upt                 0.000928   0.001916  0.484361 6.3252e-01
## accident_count     -0.112139   0.022027 -5.090872 3.3024e-05 ***
## season             -2.141691   1.197522 -1.788435 8.6338e-02 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 12.6     Adj. R2: 0.313204
##              Within R2: 0.054192
```

```
##
## Call:
## lm(formula = on_time ~ avg_rain_inch^2 + tmax + upt + accident_count +
##     route + season + PC1, data = transit_data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -70.775  -8.264   1.188  10.010  37.793
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    70.0299707  2.3968015  29.218  < 2e-16 ***
## avg_rain_inch   1.5970051  0.4181081   3.820 0.000134 ***
## tmax            0.0903018  0.0259163   3.484 0.000495 ***
## upt             0.0028468  0.0004315   6.598 4.40e-11 ***
## accident_count -0.1114160  0.0071092 -15.672  < 2e-16 ***
## route          -0.0194474  0.0007561 -25.720  < 2e-16 ***
## season         -2.6393521  0.4067065  -6.490 9.02e-11 ***
## PC1            -1.8226588  0.1232226 -14.792  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.39 on 9840 degrees of freedom
## Multiple R-squared:  0.1129, Adjusted R-squared:  0.1123
## F-statistic: 178.9 on 7 and 9840 DF,  p-value: < 2.2e-16
```