Null_Data_Challenge

Roger Yuan

9/3/2019

Import Datasets

Revenue for Each Driver

```
## # A tibble: 937 x 2
##    driver_id                        total_rev
##    <chr>                                <dbl>
##  1 5ccc0e6dc9c7475caf785cdce7b8eb7a    12350.
##  2 3788dc9e91f1548816ce8b5af07ddadc    12305.
##  3 4eb382d1f7d50fae1294964263d1ce82    10782.
##  4 6b65c06851e944351dd285a1eb729499    10709.
##  5 844e9be5a30d8d9c1f8e9ddb086ff717    10614.
##  6 c07499b5a6f1090f2fb263ec6ac0660c    10013.
##  7 af452fac966efa9d089953c99045bb20     9712.
##  8 689bdf87fb2de49f98bf4946cfaa5068     9688.
##  9 55bec90600d21bd3513366d218f2b2f2     9407.
## 10 abc63585621a8cc49099bc9dde677f27     9360.
## # ... with 927 more rows
```

Determining the Time of Day for Rides

Driver's Lifetime

```
## # A tibble: 837 x 4
##    driver_id            driver_onboard_date latest_time         life_time
##    <chr>                <dttm>              <dttm>              <drtn>
##  1 75ff47d4ba4bd4480629~ 2016-03-28 00:00:00 2016-06-26 18:55:28 90.78852 ~
##  2 72ca99bb6667024a23e1~ 2016-03-28 00:00:00 2016-06-26 16:17:15 90.67865 ~
##  3 8dbfef11a650dd9658ca~ 2016-03-29 00:00:00 2016-06-27 00:05:30 90.00382 ~
##  4 479c3dccc06056867dd1~ 2016-03-28 00:00:00 2016-06-25 23:42:55 89.98814 ~
##  5 15e0ade28a7d24026b2f~ 2016-03-29 00:00:00 2016-06-26 22:38:07 89.94314 ~
##  6 7c27405cefee2fad79a8~ 2016-03-29 00:00:00 2016-06-26 20:41:08 89.86190 ~
##  7 630559dc85053c746bea~ 2016-03-29 00:00:00 2016-06-26 18:24:47 89.76721 ~
##  8 9c5a3250d9f6e6537b3b~ 2016-03-29 00:00:00 2016-06-26 17:59:52 89.74991 ~
##  9 956942174fc793c4bfb6~ 2016-03-28 00:00:00 2016-06-25 15:43:18 89.65507 ~
## 10 c1d662bc81ade6c9a1ec~ 2016-03-29 00:00:00 2016-06-26 13:59:44 89.58315 ~
## # ... with 827 more rows
```

```
## Time difference of 55.75609 days
```

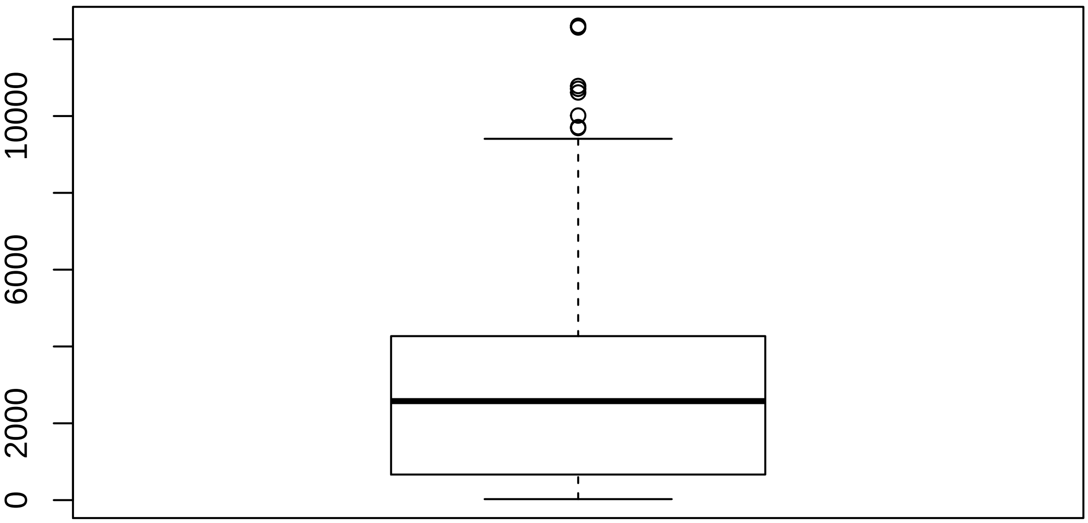
Time Between Rides

Summarizing Results

```
## # A tibble: 10 x 8
##    driver_id Total_Rev life_time Total_distance Total_duration
##    <chr>         <dbl> <drtn>             <dbl>          <dbl>
##  1 5ccc0e6d~    12350. 82.82890~        6353153         748384
##  2 3788dc9e~    12305. 61.11625~        6170230         779797
##  3 4eb382d1~    10782. 81.90313~        5056207         693727
##  4 6b65c068~    10709. 49.99959~        5089164         629203
##  5 844e9be5~    10614. 72.41716~        5004424         736296
##  6 c07499b5~    10013. 86.73431~        6179913         600682
##  7 af452fac~     9712. 65.87589~        4980036         636973
##  8 689bdf87~     9688. 84.00549~        4875955         579776
##  9 55bec906~     9407. 48.02179~        4516137         560299
## 10 abc63585~     9360. 55.99272~        4791054         569194
## # ... with 3 more variables: Total_prime_time <dbl>, Rush <chr>,
## #   Full_time <int>
```

Investigating Timestamps of Outlier Drivers

Investigating Revenues of Outlier Drivers



```
## # A tibble: 6,249 x 5
##    driver_id             ride_id              totaltime norider percentnorev
##    <chr>                 <chr>                    <dbl>   <dbl>        <dbl>
##  1 3788dc9e91f1548816ce~ 0004a3db0fb83d9900~       1.58    1.22       0.768
##  2 3788dc9e91f1548816ce~ 004d182a6d22d9c220~      34.5     6.52       0.189
##  3 3788dc9e91f1548816ce~ 007d84776ff876e504~       7.58    2.95       0.389
##  4 3788dc9e91f1548816ce~ 00b8de8efce6ac756f~      36.4     1.43       0.0394
##  5 3788dc9e91f1548816ce~ 0111b767b1ab732964~      20.3     2.22       0.109
##  6 3788dc9e91f1548816ce~ 016cd8d06dfad03d63~      40.1     2.78       0.0694
##  7 3788dc9e91f1548816ce~ 01fd89ca0a024a360a~      10.2     2.02       0.197
##  8 3788dc9e91f1548816ce~ 02026b7278ec4745aa~      17.0     1.78       0.105
##  9 3788dc9e91f1548816ce~ 021b8a875091731b88~       9.77    1.18       0.121
## 10 3788dc9e91f1548816ce~ 023abb75ac4eba8557~      10.3     1.78       0.173
## # ... with 6,239 more rows
```

```
##                             driver_id totalprimetime
## 218 3788dc9e91f1548816ce8b5af07ddadc          18100
## 309 4eb382d1f7d50fae1294964263d1ce82          17475
## 355 5ccc0e6dc9c7475caf785cdce7b8eb7a          16650
## 402 689bdf87fb2de49f98bf4946cfaa5068          12375
## 413 6b65c06851e944351dd285a1eb729499          19675
## 510 844e9be5a30d8d9c1f8e9ddb086ff717          13750
## 665 af452fac966efa9d089953c99045bb20          11125
## 723 c07499b5a6f1090f2fb263ec6ac0660c           8050
```

```
##                            driver_id totalprimetime
## 413 6b65c06851e944351dd285a1eb729499          19675
## 218 3788dc9e91f1548816ce8b5af07ddadc          18100
## 689 b6ec72d2f14dcc90a4e7fd25bd12e9a7          17650
## 309 4eb382d1f7d50fae1294964263d1ce82          17475
## 330 55bec90600d21bd3513366d218f2b2f2          17400
## 355 5ccc0e6dc9c7475caf785cdce7b8eb7a          16650
## 323 531a726b5b0c925a1aa24b5a9d5ac333          15750
## 762 cf27028c2fe4a9fe00795d0d4dd23a18          15525
```

Regression Analysis

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```
## Loading required package: foreach
```

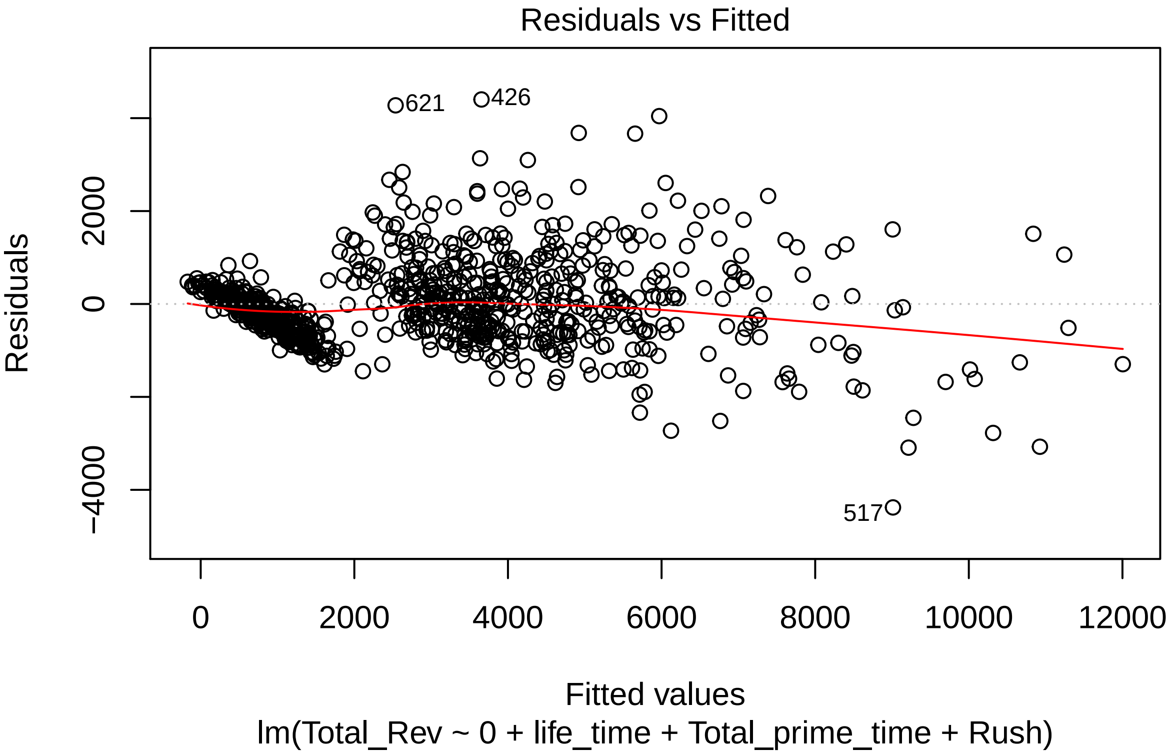
```
##
## Attaching package: 'foreach'
```

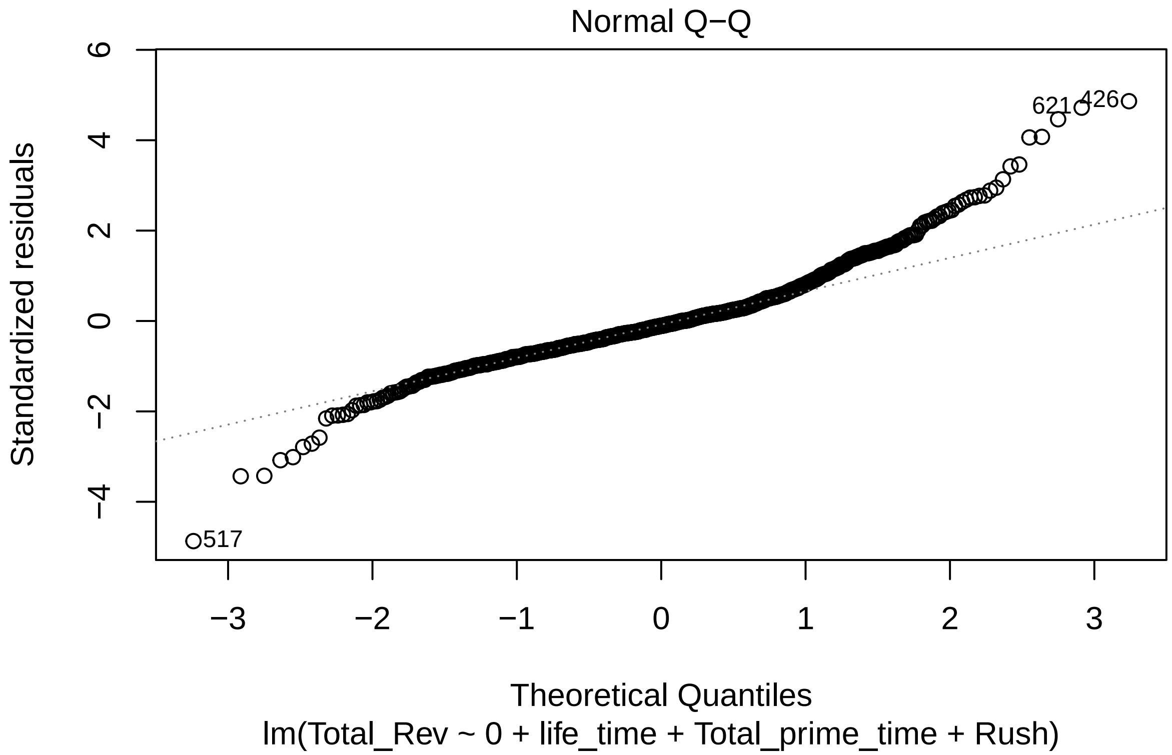
```
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
```

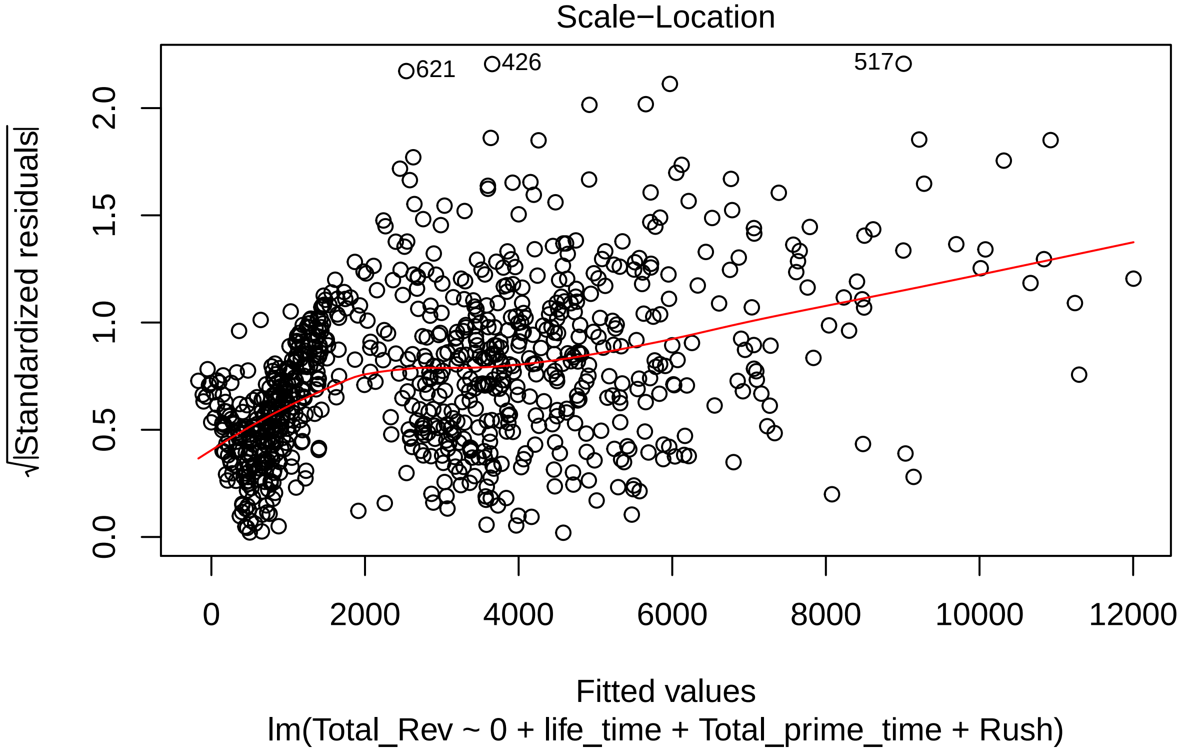
```
## Loaded glmnet 2.0-18
```

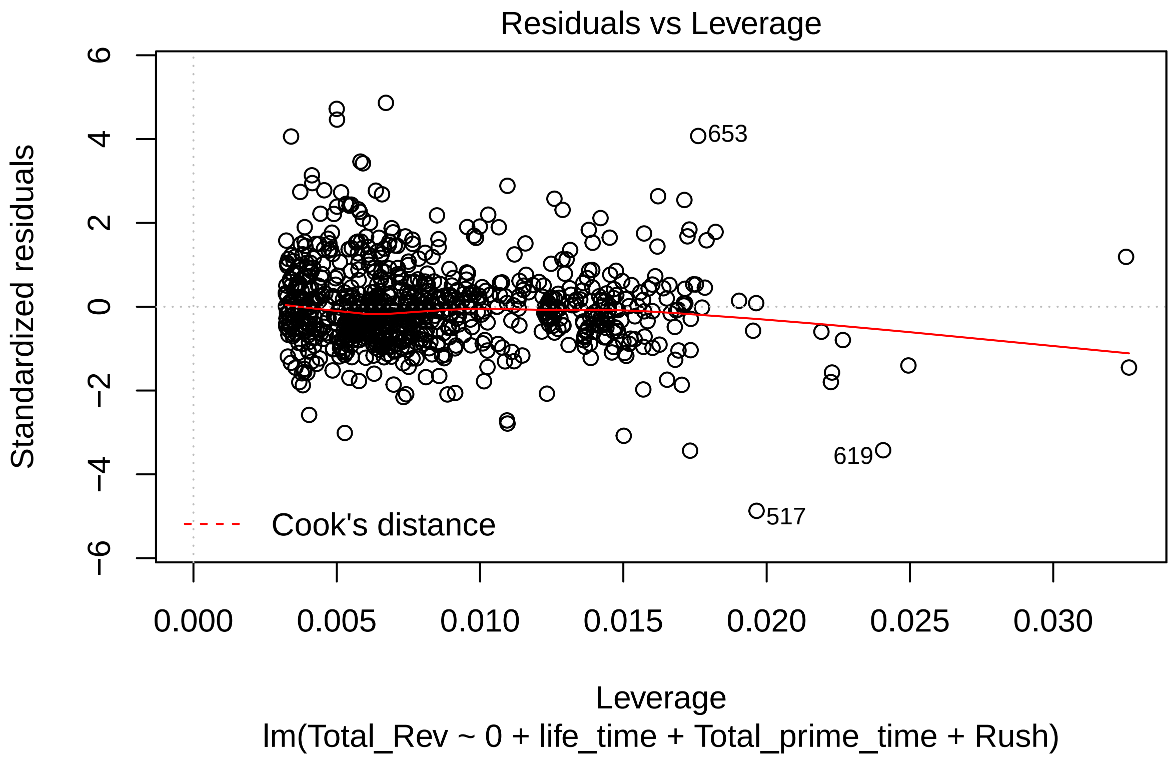
```
##      row col
## [1,] 425   3
```

```
##
## Call:
## lm(formula = Total_Rev ~ 0 + life_time + Total_prime_time + Rush,
##     data = Driver_summary[, -1])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4378.5  -521.6   -97.4   377.6  4401.7
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## life_time          17.41249    1.60660  10.838   <2e-16 ***
## Total_prime_time    0.57420    0.01042  55.093   <2e-16 ***
## Rushafternoon    -107.13456  129.56611  -0.827   0.4085
## Rushevening rush -147.69446   99.46064  -1.485   0.1379
## Rushmorning      -386.96812  125.39059  -3.086   0.0021 **
## Rushmorning rush -214.85456  107.72997  -1.994   0.0464 *
## Rushnight        -163.12887  106.73382  -1.528   0.1268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 908 on 830 degrees of freedom
## Multiple R-squared:  0.9444, Adjusted R-squared:  0.944
## F-statistic:  2015 on 7 and 830 DF,  p-value: < 2.2e-16
```




Normal Q−Q
Standardized residuals
6
4
2
0
−2
−4
621
426
517
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Total_Rev ~ 0 + life_time + Total_prime_time + Rush)


Scale−Location
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
621
426
517
0
2000
4000
6000
8000
10000
12000
Fitted values
lm(Total_Rev ~ 0 + life_time + Total_prime_time + Rush)



```
##        life_time Total_prime_time    Rushafternoon Rushevening rush
##       17.4124882        0.5741959    -107.1345622     -147.6944556
##      Rushmorning Rushmorning rush        Rushnight
##     -386.9681190     -214.8545569    -163.1288684
```

```
## [1] 837
```

```
##   (Intercept) life_time Total_distance Total_duration Total_prime_time
## 1    23.63603 0.1080286   0.0007412914    0.007395187        0.1081479
##   Rushevening rush Rushmorning Rushmorning rush Rushnight Full_time
## 1                0           0                0  11.02732         0
```

As seen from regression analysis, total prime time and life time of a driver play key roles in determining total revenue for a driver. Whether a driver is classified as full time or part time does not matter so much since the main variations variables Life Time, Total Prime Time, and Rush Hour capture most of the information in the data.

```
mean(ride_ids$revenue[which(ride_ids$dayinfo=='night')])
```

```
## [1] 13.09894
```

```
mean(driver_lifetime$life_time, na.rm=T)
```

```
## Time difference of 55.75609 days
```

```
lifetimesummary <- boxplot.stats(as.numeric(driver_lifetime$life_time))
lifetimesummary
```

```
## $stats
## [1]  1.74000 42.84226 57.96610 73.11388 90.78852
##
```

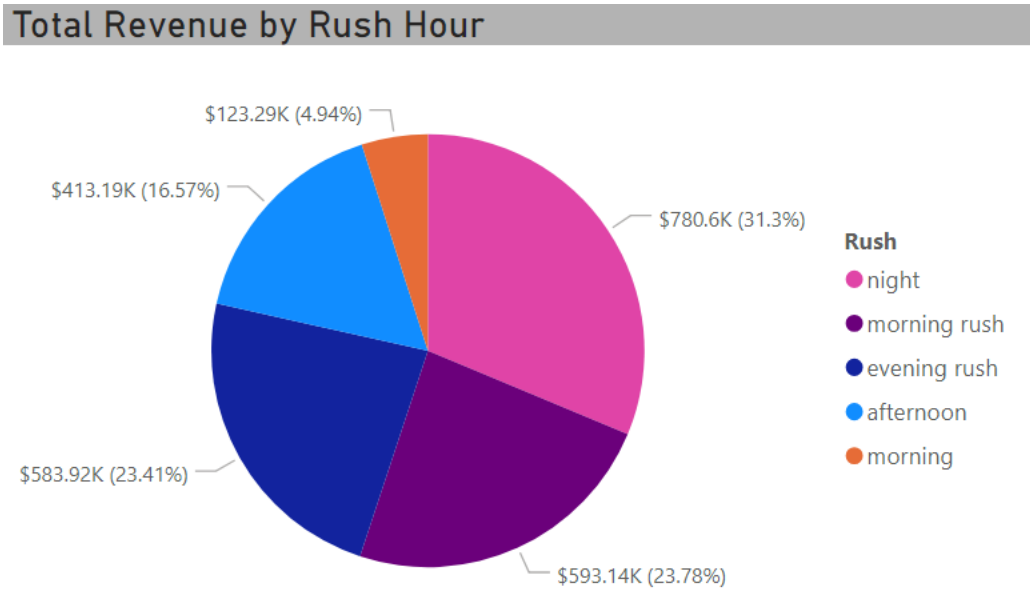


Figure 1: The pie chart suggests that drivers working during night and rush hours generate the most revenue.

```
## $n
## [1] 836
##
## $conf
## [1] 56.31189 59.62031
##
## $out
## numeric(0)
```

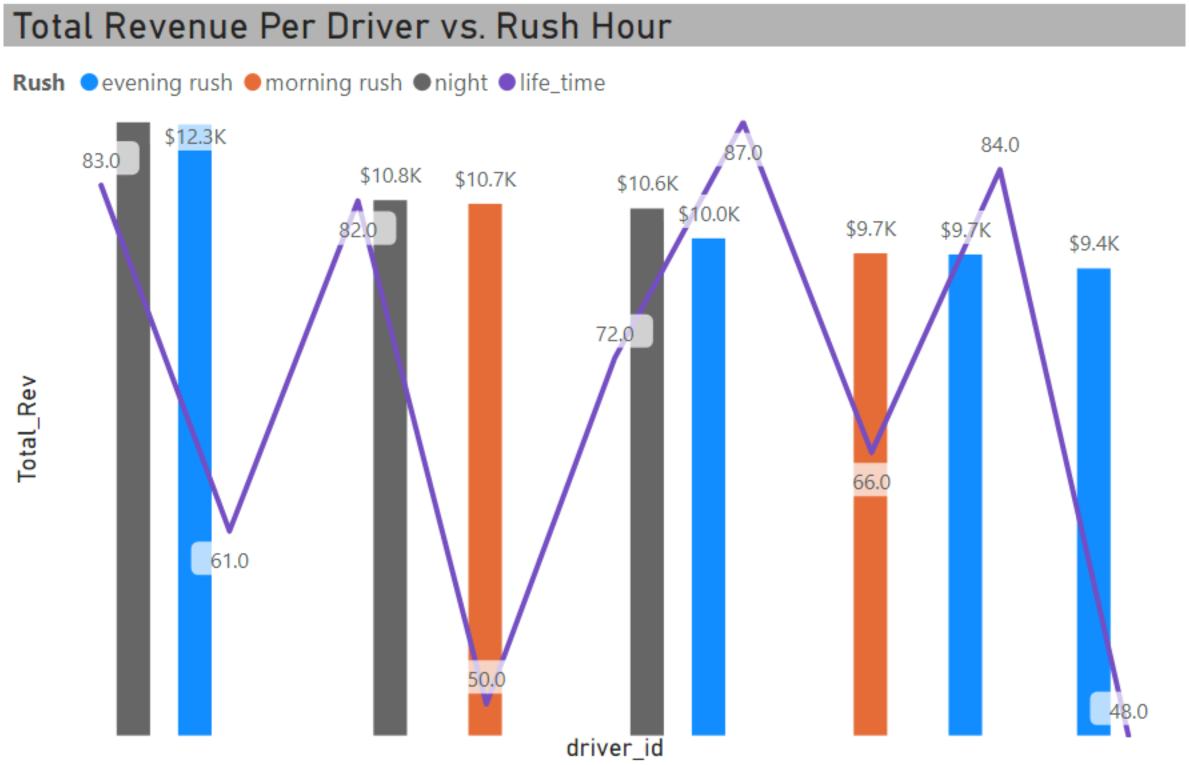


Figure 2: The column and line chart reveals that lifetime (total work days) isn't the only important factor in determining a driver's revenue. The time of the day a driver works at also plays a key role in evaluating a driver's lifetime value to Lyft.

A driver's lifetime value would be calculated using a linear combination of projected lifetime revenue, average prime time bonus per ride, and proportion of rides driven during rush hour or late nights. After normalizing each component, we multiply them together to get a lifetime value.

Revenue is often an important indicator of an employee's worth to a company. At the end of the day, an employee's contribution to their company's bottom line determines how important they are to the company, at least financially. To determine a driver's financial value, we found their average daily revenue by totaling the revenue of their rides during their lifetime and dividing by their lifetime before multiplying by the average driver lifetime. This revenue calculation accounts for prime time bonuses as well.

However, while revenue is important, it is not all a driver has to offer to Lyft. We believe that a driver who works for Lyft during times of high demand, such as holidays or rush hours, would be of much value to the company. Not only would they bring in more revenue from the prime time bonuses that have taken effect, but they also bolster Lyft's consistency and dependency during times when people rely on them the most. Moreover, they increase Lyft's ability to contend against competitors when there are more potential riders unfamiliar with the service or are not loyal to a specific one. By allowing Lyft to provide its services to more riders during especially busy, critical times, these drivers have increased worth towards Lyft that cannot be described by revenue alone.

In order to represent this, we summed up all the prime time bonuses for each driver as an approximate measure of how often they drive during high demands. Larger total prime time bonuses indicate a tendency to work during busy times and lower totals indicate otherwise. We once again normalize this factor.

Another factor we wish to consider is how often a driver works during rush hours. This is somewhat similar to the previously mentioned total prime time, but the number of rides during rush hour or late nights represents a more consistent work schedule. Because riders who hail rides during these times often need to, often due to work, we believe Lyft drivers who work within these hours are not only fulfilling demand but also building customer loyalty. Because many regularly require rides during these times, they tend to stick with a ride hailing company, and drivers who help saturate this market with Lyft options not only capitalize on revenue but also aid in creating loyal, long-term customers for Lyft. Due to their contributions towards brand marketing and loyalty, these drivers are also very valuable to Lyft.

To get this information we labelled each ride with a time of day depending on when the ride was requested. We then summed up the number of rides that occured during rush hours (5-10am and 3-7pm) and once again normalized it.

Driver Projected Lifetime

The average projected lifetime of a driver is approximately 55.75 days, or a little under 2 months based on the data given. Furthermore, the minimum lifetime is 1.7 days, the first quartile is 43 days, the median is 58 days, the third quartile is 73 days, and the maximum is 91 days. The distribution of lifetimes is only slightly left skewed. We calculated the lifetime of a driver by subtracting their onboard date from their last recorded ride date. However, because the data was only representative of a 3 month period the projected lifetime in this dataset is likely shorter than the true projected lifetime of a Lyft driver.

```r
boxplot(as.numeric(driver_lifetime$life_time), main="Driver Lifetime in Days", ylab="Lifetime")
```

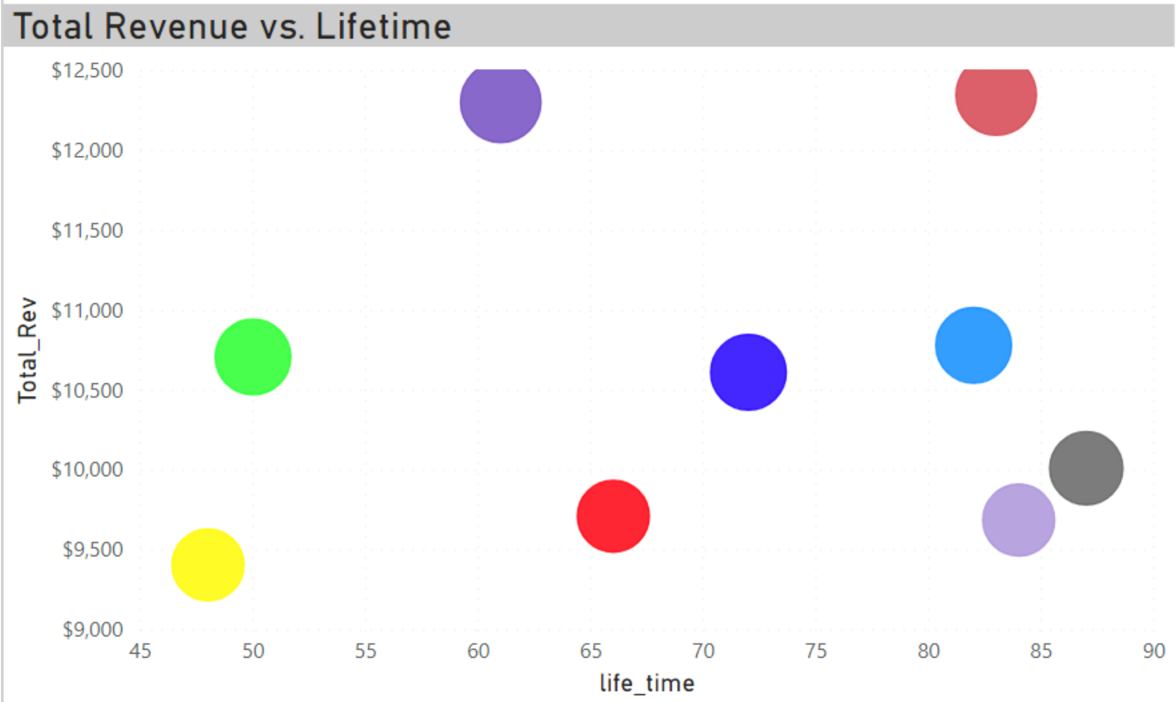
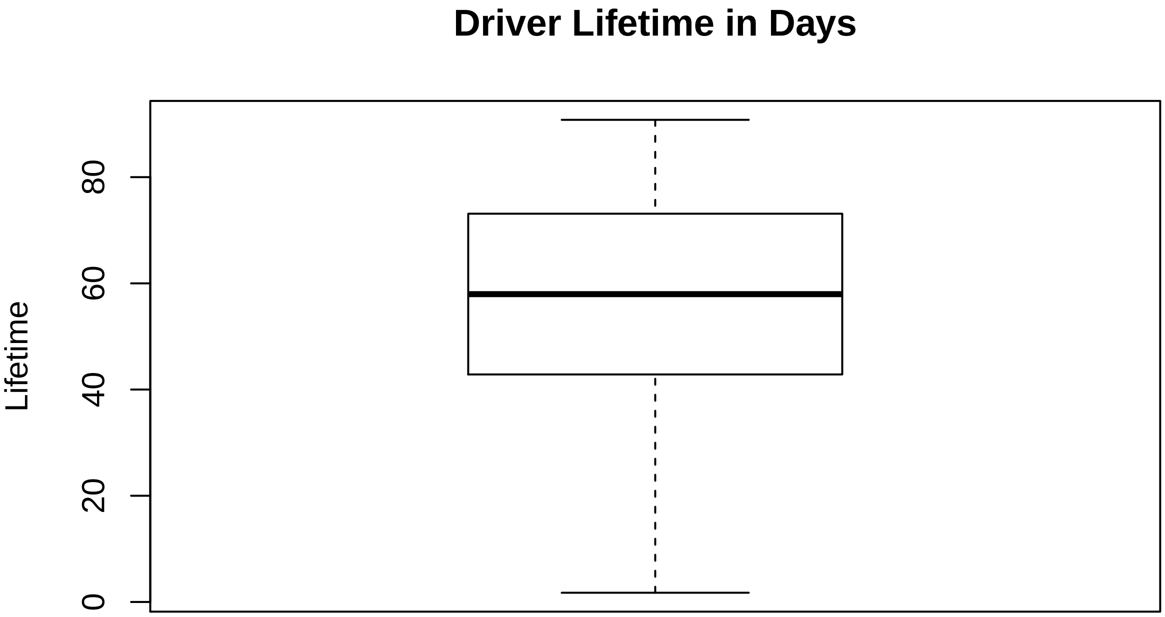


Figure 3: The bubble chart suggests little correlation between the total revenue and lifetime of a driver, refuting the claims of a multiple regression model.



As expected, the lifetime of a driver has no relation to the total revenue of a driver. The correlation between the two is 0.52, which can be seen by plotting the lifetime vs the total revenue. This makes sense, as a part-time driver working the same amount of time as a full-time driver will make much less. Thus, we can rule out lifetime as a factor in a driver's lifetime value.