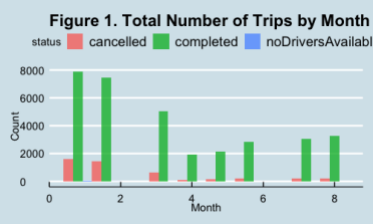
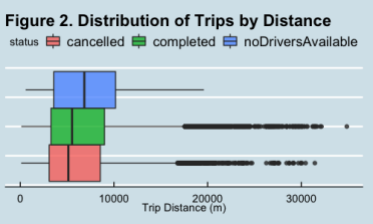
Minnesota Viking Transit (MVT) Cancelation Analysis Report

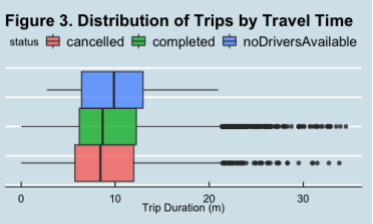
07/12/2021

Overview

There were around 38k total trips (both canceled and completed) between January and August 2020 scheduled using Minnesota Viking Transit (MVT). Trips were geographically clustered. Almost 50% of all trips within this time period occurred in the months of January and February. The number of trips for both canceled and completed dropped drastically in April (fig 1); however, this is expected since COVID-19 lockdown measures were introduced in mid-March for the state of Minnesota. 75% of trips were under 10km, almost all trips were under 20km (fig 2). The median trip duration is around 10 minutes (fig 3).







On average, 11.9% (4552) trips were canceled, with January and February having a higher cancelation rate of 16.5%. The majority of trips were canceled by end user, as opposed to no drivers available.

Method

Two regression models were built in order to understand what groups of riders are canceling more often and why. The first model followed a binomial distribution with logit link to quantify the statistical significance of trip duration, end point type, and travel month. The variable trip distance was left out due to its correlation with trip duration.

The second model followed a Poisson distribution to examine the effect of population size, median household income, proportion of nonwhite population, and proportion of zero car households on the probability of a trip cancelation. This model includes an offset for total number of trips per census block group since this variable is dependent for the number of cancelations. In order to run this model, Haversine calculations were used to assign the nearest census block group centroid to the requested pick-up location of each trip record. Only 85 of 4111 census block groups were associated with trip records, meaning the user base of MVT is clustered.

Findings

The groups of riders are classified as two categories: characteristics of the trips themselves and the socioeconomic factors of rider demographics.

Firstly, for the trips themselves, the pandemic restrictions have a significant effect on the probability of cancelations starting in the month of March. On average, trips occurred during the pandemic months (March-August) has an average of 0.28% higher likelihood of being canceled than pre-pandemic. As for end point types, scheduled trips are most likely to be canceled and 'next available' trips are least likely to be canceled. Last but not least, a 5-minute increase in expected trip duration will result in trips being 2.5% more likely to be canceled.

As for the effect of demographic factors, the analysis reveals that percentage of zero-car households, median household income, and proportion of non-white population are all significant in determining the likelihood of a trip cancelation. This is based on the assumption that the census block group the starting trip location is in is where the user resides. A 1% increase in each variable will result in 1 more trip cancelation in the area.

Recommendations

Overall, longer expected trip durations, trips scheduled during the pandemic months, a higher proportion of non-white population, a higher percentage of zero-car households, and a higher median household income make up the group with high cancelation tendencies.

Since rider cancelation tendency is highly correlated with both socioeconomic factors and trips characteristics, strategic marketing is recommended targeting riders in these categories. Educational-purpose push notifications can be sent to users who schedule long trips, frequently start trips in census block groups with high proportion of non-white population, high income, and high percentage of zero-car households, as well as have a record of cancelations. These targeted notifications can act as a reminder for proper trips etiquette. Additional incentives such as ride discount could also be provided for these users to reduce overall cancelation rate.

Next steps of analysis could utilize shapefile of census block groups in Minnesota and further explore origin destination of trips departed from high tendency areas.

Appendix Code

Appendix Code

Load data

```
# load data
trips <- readxl::read_excel('SampleDataset_Jan2020_Aug2020.xlsx')
census <- readxl::read_excel('SampleDataset_Census_Minnesota.xlsx')

# add month info
trips$travelMonth <- as.numeric(substr(trips$requestedPickupTsTimestamp, 6, 7))
```

Exploratory Analysis

```
# check number of trips
nrow(trips) #38,250 total trips

# check trip distribution over time
hist(trips$travelMonth)
nrow(trips[which(trips$travelMonth == 1),]) #9520 january
nrow(trips[which(trips$travelMonth == 2),]) #8909 feburary
#january and feburary alone makes up almost 50% (48%) of all trips within this time period

# check number of canceled trips
nrow(trips[which(trips$status == 'cancelled'),]) #4552 canceled trips
nrow(trips[which(trips$status == 'cancelled'),])/nrow(trips) #11.9% trips canceled

# cancel trips distribution over time
hist(trips[which(trips$status == 'cancelled'),]$travelMonth)
nrow(trips[which(trips$status == 'cancelled' & trips$travelMonth == 1),]) #1580 january
nrow(trips[which(trips$status == 'cancelled' & trips$travelMonth == 2),]) #1462 feburary
(1580+1462)/(9520+8909) #16.5% cancelation in january and feburary

# visualizations
#hist(trips$travelDistance)
ggplot(trips, aes(travelMonth, fill = status)) +
  theme_economist() +
  ggtitle('Figure 1. Total Number of Trips by Month (Jan - Aug 2020)') +
   geom_histogram(alpha = 0.8, position = 'dodge', binwidth = 0.8)  + #or position identity
  xlab('Month') + ylab('Count')

trips$minutes <- trips$travelDuration/60
```

```r
ggplot(trips, aes(travelDistance, fill = status)) +
  theme_economist() +
  ggtitle('Figure 2. Distribution of Trips by Distance') +
   geom_boxplot(alpha = 0.8, position = 'dodge')  + #or position identity
  xlab('Trip Distance (m)') + ylab('Count')


ggplot(trips, aes(minutes, fill = status)) +
  theme_economist() +
  ggtitle('Figure 3. Distribution of Trips by Travel Time') +
   geom_boxplot(alpha = 0.8, position = 'dodge')  + #or position identity
  xlab('Trip Duration (m)') + ylab('Count')
```

Haversine Distance Calculation

```r
# change from degree to radian
trips$lat <- trips$requestedPickupLatitude*pi/180
trips$lon <- trips$requestedPickupLongitude*pi/180
census$lat <- census$Latitude_Centroid*pi/180
census$lon <- census$Longitude_Centroid*pi/180

# haversine calculations
distance.hf <- function(lon1, lat1, lon2, lat2) {
  R <- 6371 # Earth mean radius [km]
  delta.lon <- (lon2 - lon1)
  delta.lat <- (lat2 - lat1)
  a <- sin(delta.lat/2)^2 + cos(lat1) * cos(lat2) * sin(delta.lon/2)^2
  c <- 2 * asin(min(1,sqrt(a)))
  d = R * c
  return(d) # Distance in km
}

dat <- data.frame()
# loop through and attach census information
for (i in 1:nrow(trips)) {
  temp <- data.frame()
  for (j in 1:nrow(census)) {
    # create distance matrix
    temp <- rbind(temp, distance.hf(trips$lon[i], trips$lat[i], census$lon[j], census$lat[j]))
  }
  # reverse look up index and attach
  dist = min(temp)
  index = which(temp == dist)
  dat <- rbind(dat, cbind(trips[i,], census[index,], dist))
}
```

Model 1. Logistic Regression

```r
# build y variable
dat2 <- dat
dat2$cancelled <- ifelse(dat2$status == 'completed', 0, 1)
# build x variables
dat2$endpoint <- as.factor(dat2$endpoint)
```

```r
dat2$travelMonth <- as.factor(dat2$travelMonth)
dat2$travelDuration <- dat2$travelDuration/60 #minutes
dat2$propnonwhite <- dat2$`ProportionNonWhitePopulation_%`
dat2$income <- dat2$`MedianHouseholdIncome_$`
dat2$pop <- dat2$`TotalPopulation_#`
dat2$car <- dat2$`ProportionOfZeroCarHouseholds_%`

# logistic regression
res1 <- glm(cancelled ~ endpoint + travelMonth + travelDuration,
            data = dat2, family= binomial(link = 'logit'))
summary(res1)

# result table
as.data.frame(summary(res1)[12])
logOddsMat = res1$coefficients
oddsMat = exp(logOddsMat)
oddsMat = oddsMat/(1 + oddsMat)
oddsMat = as.data.frame(oddsMat)
oddsMat$significance = c('***', '*', '***', '', '***','***','***','***','***','***','***')
colnames(oddsMat) <- c('Coefficient', 'Significance')
table1 <- oddsMat %>%
  kableExtra::kbl(
    caption = "Figure 1. Logistic Regression Odds Ratios Table",
    booktabs = T, digits = 3) %>%
  kableExtra::kable_classic(full_width = F, html_font = "Cambria") %>%
  kableExtra::kable_styling(latex_options = c("HOLD_position", "striped"))
table1
```

Figure 1. Logistic Regression Odds Ratios Table

	Coefficient	Significance
(Intercept)	0.109	***
endpointnextAvailable	0.657	*
endpointscheduled	0.896	***
travelMonth2	0.487	
travelMonth3	0.374	***
travelMonth4	0.237	***
travelMonth5	0.251	***
travelMonth6	0.242	***
travelMonth7	0.241	***
travelMonth8	0.243	***
travelDuration	0.496	***

Model 2. Poisson Model with Offset Variable

```
# poisson regression with offset
dat2$education <- dat2$`ProportionHighSchoolEducatedOrLess_%`
dat3 <- dat2 %>%
  select(status, `CensusBlockGroup_#`, propnonwhite, income, pop, car) %>%
  group_by(`CensusBlockGroup_#`) %>%
  summarize(total_trips = n(), canceled = sum(status!='completed'), propnonwhite = min(propnonwhite),
            income = min(income), pop = min(pop), car = min(car))

res2 <- glm(canceled ~ propnonwhite + car + income + pop + offset(log(total_trips)),
            data = dat3, family = poisson(link = 'log'))
summary(res2)

# expected counts
logOddsMat2 = res2$coefficients
oddsMat2 = exp(logOddsMat2)
oddsMat2 #expected number of canceled trips
oddsMat2 = as.data.frame(oddsMat2)
oddsMat2$Significance = c('***', '***', '**', '***', '')
colnames(oddsMat2) <- c('Expected Count', 'Significance')

table2 <- oddsMat2 %>%
  kableExtra::kbl(
    caption = "Figure 2. Poisson Regression Expected # of Cancellations",
```



```
    booktabs = T, digits = 3) %>%
  kableExtra::kable_classic(full_width = F, html_font = "Cambria") %>%
  kableExtra::kable_styling(latex_options = c("HOLD_position", "striped"))
table2
```

Figure 2. Poisson Regression Expected # of Cancellations

	Expected Count	Significance
(Intercept)	0.073	***
propnonwhite	1.007	***
car	1.008	**
income	1.000	***
pop	1.000	