Documentation for Individual Project

Purpose

The purpose of this project was to create visualization based on the data about the taxi trips that took place in Chicago in 2016.

Data description

The data had 20 variables originally. The following 9 have been primarily used, even though other variables have been sometimes used to demarcate things:-

Name	Type	Range	Example
Trip ID Number	String	Any combination of letters and numbers	85
Start and End Time	String “MM/DD/YYYY HH:MM:SS AM/PM” (Time is rounded to the nearest 15 minutes)	01/01/2016-12/31/2016 00:00:00 - 23:59:59	01/13/2016 6:15:00 AM
Length of Trip (miles and seconds)	Number	0-100 miles 0-10000 seconds	1.3 miles 100 seconds
Pickup and Drop off Location	Number, (lat, long) by community areas		610, 509
Fare	Number (in dollars)	0 - 1000	44.25
Tips	Number (in dollars)	0 - 1000	8.95
Tolls	Number (in dollars)	0-1000	2.0
Extras	Number (in dollars)	0-1000	2.5
Trip total	Number (in dollars)	0-10000	10.75
Payment Type	String	“Cash”, "Credit Card”, “Unknown”	Credit Card

Taxi company	String	any	104

Data Collection

The data was sourced from the Kaggle website, the link to which is https://www.kaggle.com/datasets/chicago/chicago-taxi-rides-2016?select=data_dictionary.csv

The data was created by the City Council of Chicago, and the process is highlight here https://design.chicago.gov/

It mentions that simply because of the way the trips are reported, not all trips have been reported but most of them have been.

Target Users

The target users for this app could be multiple. Some examples are :-

1. Taxi drivers in Chicago, who would want to know about optimum routes.
2. The city council of Chicago, for knowledge on public transport
3. Data Analysis experts and students

Target Questions

This program aims to answer several questions, primarily through visualizations. Some of those are:-

1. Average total number of trips for a given hour in the day.
2. Number of trips taking place each month in 2016
3. Trips taking place between particular locations

Data Insights

We receive the follow important insights from the data :-

1. 6-7 PM are the busiest hours for taxis in Chicago.
2. The highest number of cabs are taken on the 18th-19th of every month.

3. Most days have roughly the same number of cab count, except 31st because not every month has 31st.
4. Most number of cabs are taken in March, least in December, which could be because of holiday season.

Future Improvement Goals

The data could be improved in a number of ways. Some of those are:-

1. The processing time is incredibly slow due to the large amount of data. The data could be compressed to help with this aspect.
2. Some additional visualizations could be added.

Sources

The following sources have been used:-

1. https://data-flair.training/blogs/r-data-science-project-uber-data-analysis/
2. https://stackoverflow.com/questions/43470094/plotting-by-ggplot-in-r
3. https://ggplot2.tidyverse.org/reference/
4. https://www.r-project.org/help.html
5. https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/help

Data Tidying

The data has been tidied in the following ways :-

1. Because the project is heavily inspired from another similar project that dealt with Uber rides in NYC, the variable names have been changed accordingly.
2. Rows with empty values have been removed.