UNIVERSITY OF MAURITIUS

Faculty of Social Sciences & Humanities (FSSH)

Department of Economics & Statistics

Course

MSc Data Analytics

Year 1 Semester 1

Module

Data Handling and Analysis

STAT 5129

Final Group Coursework

Submitted to:

Mr. R. Thoplan

Submitted by:

Mavish GAJADHUR	*2521826*
Parvesh GHOORA	*2522774*
Homeswaree JOWAHEER	*2522007*
Saveena KOWLESSUR	*2521531*
Shradha NUCCA GOVEDO	*2520562*

Submission Date: 31st October 2025

NYC Yellow Trip Taxi Data Analysis 2024

Overview

This project provides a brief analysis of New York City's Yellow Taxi Trip dataset for 2024. The purpose of the project is to demonstrate reading a Parquet file into R statistical software, filtering and saving a cleaned version using the “arrow” package.

The analysis aims to uncover travel patterns, trip duration, fare distributions by payment type, and peak hours trends through visualizations and statistical analysis.

1. **Reading the Parquet file**

Download Data from URL

https://d37ci6vzurychx.cloudfront.net/trip-data/yellow_tripdata_2024-01.parquet

The dataset is public and contains no personal information;

Data is used solely for assignment purposes under fair-use conditions.

2. **Setup and Libraries**

The following packages are used for analysis

Arrow: for efficient reading and writing of Parquet files

dplyr: for data manipulation and summarization

ggplot2: for visualization

lubridate: for extracting time components from timestamps

readr: for exporting results to Comma-Separated Values (CSV)

3. **Extract data, transform and Inspection**

Set working directory and load data. The data size (in memory) is 361.9 MB and the loading time is approximately 0.75 seconds.

Perform quick checks for data structure and type.

The NYC yellow taxi dataset contained 2,964,624 trip records with the following 19 key features:

Table 1: Data description

Feature	Data_type	Description
VendorID	Discrete	A code indicating the TPEP provider that provided the record, 1 and 2
tpep_pickup_datetime	Datetime	The data and time when the meter was engaged, e.g. 2003-01-01 02:59:39
tpep_dropoff_time	Datetime	The data and time when the meter was engaged, e.g. 2003-01-01 03:05:41
passenger count	Discrete	The number of passengers in the vehicle.
trip_distance	Continuous	The elapsed trip distance in miles reported by the taximeter.
RatecodeID	Discrete	The final rate code in effect at the end of the trip. 1 = Standard rate 2 = JFK 3 = Newark 4 = Nassau or Westchester 5 = Negotiated fare 6 = Group ride
store_and_fwd_flag	Character	This flag indicating whether the trip recorded was held in vehicle memory before sending to the vendor, aka "store and forward", because the vehicle did not have a connection to the server. Y = store and forward trip N = not a store and forward trip
PULocationID	Discrete	TLC Taxi Zone in which the taximeter was engaged
DOLocationID	Discrete	TLC Taxi Zone in which the taximeter was disengaged
payment_type	Discrete	A numeric code signifying how the passenger paid for the trip. 1 = Credit card 2 = Cash 3 = No charge 4 = Dispute 5 = Unknown 6 = Voided trip
fare_amount	Continuous	The time-and-distance fare calculated by the meter.
extra	Continuous	Miscellaneous extras and surcharges. Current, this only includes the $0.50 and $1 rush hour and overnight charges.
mta_tax	Continuous	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
tip_amount	Continuous	Tip amount – This field automatically populated for credit card tips. Cash tips are not included.
tolls_amount	Continuous	Total amount of all tolls paid in trip
improvement_surcharge	Continuous	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
total_amount	Continuous	The total amount charged to passengers. Does not include cash tips.
congestion_surcharge	Continuous	Total amount collected in trip for NYS congestion surcharge.
Airport_fee	Continuous	$1.25 for pick up only

There are 700,810 missing values in the dataset, 140,162 for each of the following variables: passenger_count, RatecodeID, store_and_fwd_flag, congestion_surcharge and Airport_fee.

There are some features with negative which is not relevant. Features with negative values as well as those with missing values are not taken for analysis.

For feature 'trip distance', records are filtered and values greater than zero are retained for analysis as distance cannot be less than zero.

A new variable trip_duration is created based on tpep_pickup_datetime and tpep_dropoff_time, where trip duration in minutes is equal to dropoff time minus pickup time. Non-positive distances are filtered out.

Payment type has been converted to descriptive labels as: 1 to "Credit Card", 2 to "Cash", 3 to "No Charge", 4 to "Dispute", 5 to "Unknown" and 6 to "Voided trip", to analyse trends for payment preferences.

4. Data Analysis and visaulisations

Analysis of payment type

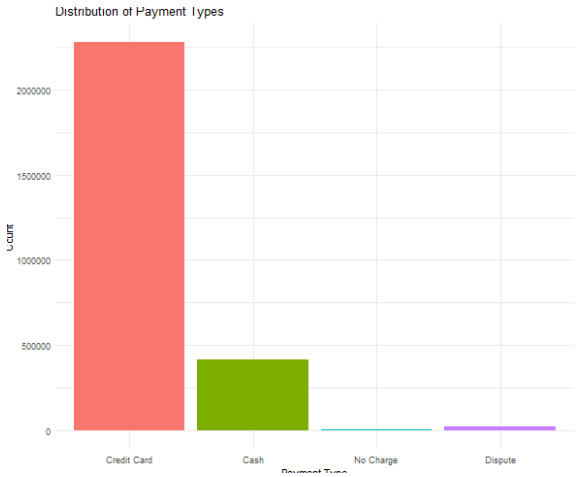


Table 2: Count by mode of payment

Payment type	Counts	Percentage
Credit Card	2,319,046	78.2
Cash	439,191	14.8

Table 2 above shows that most NYC taxi rides were paid by credit card on a total amount of 2,964,624 trips, indicating the city's trend toward cashless transactions. Credit card payments are the preferred choice among passengers, making up 78% of the total, while 15% still opt for cash paymentsamong the other types of payment.

Analysis of peak hours

Hours of the day with trip counts is shown in tables 3 and 4 below.

Table 3: Hours with highest number of taxi trips

Time of the day	Count
22	200,136
21	195,327
20	182,296
19	180,949
18	175,114

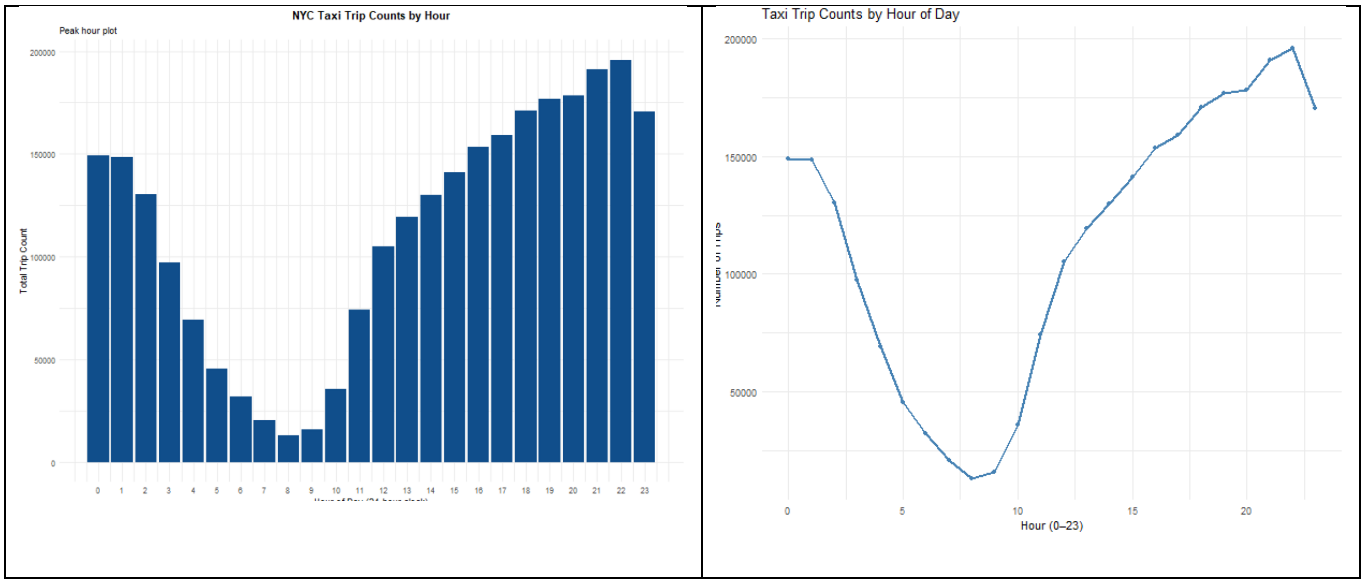
Table 4: Hours with least number of taxi trips

Time of the day	Count
10	36,398
6	32,906
7	21,232
9	16,100
8	13,266

Peak hour: 22 o'clock with 200,136 trips

Off-peak hour: 8 o'clock with 13,266 trips

Taxi usage rises sharply from late afternoon, peaking around 10 PM (22hours) and quietest hour is around 8 AM.



The plots above show a clear pattern of more taxi rides early in the morning around 1 to 2 AM, then trips level off till 8 to 9 AM, and there's another increase in the afternoon around noon to reach a peak at 10 PM. These insights help us understand how people use taxis throughout the day.

5. **Output saved**

All output results are saved for reproducibility. Scripts are organised for future use and reporting.

6. Example code snippet

```
library(arrow)
library(dplyr)
library(ggplot2)
library(lubridate)
library(readr)

# Load the dataset
taxi_df<-read_parquet("data/yellow_tripdata_2024-01.parquet")
# Basic statistics
print(taxi_df.describe())
str(taxi_df)

# Peak hour analysis
# Extracts the hour from the `tpep_pickup_datetime` column.
peak_hours_data <- taxi_df %>%
  mutate(hour = hour(tpep_pickup_datetime)) %>%
  group_by(hour) %>%
  summarise(trip_count = n()) %>%
  arrange(hour)

# Visualize fare distribution
png("outputs/peak_hour.png", width = 800, height = 600)
ggplot(peak_hours_data, aes(x = hour, y = trip_count)) +
  geom_bar(stat = "identity", fill = "dodgerblue4") +
  labs(
    title = "NYC Taxi Trip Counts by Hour",
    subtitle = "Peak hour plot",
    x = "Hour of Day (24-hour clock)",
    y = "Total Trip Count"
  ) +
  scale_x_continuous(breaks = 0:23) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```