Taxi Trips Case Study

Prepared for:

Katalyze Data

Graduate Scheme

Document History

Revision	Author	Date	Details
0.1	Barbara Mikulášová	27 Jan 2022	Initial Draft
0.2	Lucy Parkin	31 Jan 2022	Internal Review
0.3	Barbara Mikulášová	08 Feb 2023	Amendments
1.0	Lucy Parkin	22 Feb 2023	Published
1.1	Barbara Mikulášová	18 April 2024	Amendments

Table of Contents

1 Introduction

This case study should be completed using the R programming language.

A client, who is a Green Taxi provider in New York City, supplied you with some current and historical data about their taxi business, as well as the trip data of their competitors in city. In the appendix, the client enclosed a spreadsheet containing information about the different tables, their columns, and details about the collected data. Use this section to better understand the data sets you will be working with and make decisions about the data wrangling.

After the initial discussions with the client, you have been tasked to carry out analysis using only the 2021 data entries to better understand the following issues:

1. Compare the profits and the total number of rides of Green Taxi and its competitors.
2. How has the profit changed over the course of 2021 for Green Taxi and Uber companies?
3. Which trip type has brought the most revenue to the client?
4. Which pickup locations are the most popular? Are there any pickup areas where the client is underperforming significantly in comparison to the competitors?
5. The number of Green Taxi trips per weekday. Are there any days where the number of trips is significantly higher than others?
6. The peak hours for taxi services per taxi vendor
7. Which taxi provider had the most disputed trips?
8. How long does an average taxi trip last?
9. Which one of the following features has effect on the tip amount: trip length, payment type, passenger amount, drop off location or trip type?

You will be required to support any findings with appropriate graphic representations of the results.

Additionally, Green Taxi has asked you to profile any missing data and offer some suggestions on how they can improve their data reporting strategies.

2 Requirements

The following section provides information about what is expected of you during this case study.

2.1 Packages

The mandatory sections of this case study should be completed primarily using techniques seen during the training courses. The use of Base R, along with any packages from the Tidyverse is all that should be needed to develop the code for the project.

The optional Section D requires an R Markdown package to generate a pdf report with the findings.

Any statistical packages are permitted for the optional Section E.

2.2 Programming Standards

Professional programs adhere to the following principles:

1. Programs generate reproducible output:
    - Each program is free from errors, warnings and any unnecessary notes that indicate potentially inaccurate results, unless suitably justified. (Any such justifications should be made within the report).
2. Auditable and Documented:
    - Any programs created are appropriately documented with a header and comments for ease of readability.
3. Programs are organised and efficient:
    - A program should be easily executed by another person, even if the initial input data is moved to a new storage location.
    - Data, reports, programs, and other case study components should be stored in an organised folder structure.
    - The program code is clear, uses consistent naming conventions, and uses programming techniques readily understood by colleagues and peers.
    - The program should avoid unnecessary processing, combining multiple steps into fewer steps where clear and appropriate to do so.

These principles should be kept in mind while completing this assignment.

2.3 Case Study Data Sources

The following three data sets are provided for you in this case study. Note: a detailed data dictionary is provided in the appendix and should be used to fully understand the data.

- **Taxi_trips.csv** contains general information on the reported taxi trips.
- **Taxi_time_location.csv** contains information about the pickup and drop of time and location for each taxi trip.
- **Taxi_zone_lookup.csv** contains a record of locations ID codes and their corresponding New York City zones and boroughs.

3 Summary of Deliverables

This case study consists of 3 core sections, (Section A, B and C) followed by Section D that can be completed if you have attended training for R Markdown. Section E is completely optional.

3.1.1 Solutions to the Sections A-D

The completed project should be submitted as a single zip folder containing the recommended project folder structure. Relevant data and program outputs should be organised appropriately.

Someone picking up your project should be able to assign a new path to the project root folder and have any programs run without errors.

The answers to the business questions and observations about the data should be clearly formulated in the comments.

The solution for Section D should be written and stored as an R Markdown file which can be rendered into a pdf document, containing the analysis solutions. Store the report in the REPORTS folder.

3.1.2 Solution to the optional Section E

The solutions to the Section E should be stored as a separate program. The user should be able to set the path to the project folder structure once and run the program without any errors.

Any assumptions or decisions about the data, including the analysis results, should be clearly documented in the program's comments.

If your code generates any random values, such as a population sample, set a seed to ensure reproducible results.

4 Assignment

This assignment is split into 5 sections.

A: Importing Data

B: Transforming Data

C: Data Analysis and Visualisation

D: Reporting using R Markdown - *optional*

E: Hypothesis Testing - *optional*

4.1 Section A: Importing Data

Create a Project folder with the following folder structure. A DATA folder should contain two additional folders: A RAW DATA folder should contain the initial data sets, and a CLEAN DATA folder is where the transformed and clean data is exported and stored. A PROGRAMS folder should contain any R program files created. Any summary tables or output containing the results of the analysis should be stored in the REPORTS folder.

Create a program to import each of the data files and explore their structure and content. Check if all columns were assigned correct data types upon their import. Pay attention to the datetime columns and format them if necessary.

Do not forget to comment this section appropriately, including any assumptions about the data or importing related issues.

4.2 Section B: Data Wrangling

To prepare for the later analysis and reporting, transform and clean the provided data, using the following guide.

4.2.1 Data Transformation

- Using the **taxi_trips** table change following columns into character data types: **VendorID**, **RatecodeID**, **PaymentType** and **Trip_type.** Use the data dictionary found in the Appendix section for reference. Convert character type columns into factors when appropriate.

- For all tables, rename columns to reflect a consistent naming convention.
- Transpose the **taxi_time_location** table to create **taxi_time_location_wide**. Ensure it follows the principles of tidy data, with one row per unique identifier. Ensure final columns are named appropriately.
  Hint: The resulting data set should have 5 columns and 83 691 rows.
- Using the **taxi_time_location_wide** data frame from the previous step, create the following additional columns:
  - pickup_year
  - pickup_date
  - pickup_weekday
  - pickup_hour

4.2.2 Data Quality Diagnosis

Perform exploratory data analysis on all the data. Aim to identify any data quality issues such as the presence of duplicates, missing values, and extreme or unexpected values in the numerical columns. Note your observations in the code's comments.

From your exploratory analysis, apply any data cleaning steps you feel appropriate to the tables.

After cleaning the data frames, save and export them as csv files into the CLEAN DATA folder.

4.2.3 Missing Data Analysis

Identify and investigate missing data in each table to advise the client on how to improve their data collection practices. Export a summary of missing data from each table, into the REPORTS folder.

After identifying the missing values in each data frame, decide on the appropriate next steps to handle these missing values based on your existing knowledge. Do not forget to record your reasoning in the comment section for transparency and reproducibility of your work.

4.2.4 Preparing the Data for Profiling and Analytics

Merge the cleaned tables **taxi_trips** and **taxi_location_time_wide** to create a single table named **taxi_trips_nyc.** It should contain only observations from the year 2021.

Also create a join to **taxi_zone_lookup** so that the location information can be included for any pickup recorded. Rename the **zone**, **borough,** and **service_zone** columns to **pickup_zone**, **pickup_borough**, and **picklup_service_zone**.

Save, and export the table as a csv file into the CLEAN DATA folder.

4.3 Section C: Data Analysis, Visualisation, and Reporting

To answer the following business questions, create summary tables and save the results to the REPORTS folder. Analyze the output and write your answers in the comments underneath the relevant sections:

1. Find the average amount of money spent on taxi trips with the Green Taxi vendor across the week. Are there any weekdays where the number is significantly higher than others?
2. Do any of the following features have an effect on the tip amount for Green Taxi? Hint: use a scatter plot to investigate the relationship between the trip distance and the tip amount.
    - Pickup hour
    - Payment type
    - Passenger count
    - Pick up borough
    - Trip distance
3. Which taxi vendor accumulated the most profit in 2021?
4. What was the total number of rides, excluding voided trips, of Green Taxi and its competitors?

Using appropriate graphs, analyze the output and answer the following questions:

5. How has the profit amount changed over 2021 for both the Green Taxi and Uber companies. Plot both series on the same graph.
6. Which trip type has brought the most revenue to Green Taxi?
7. Which pickup borough location is the most popular for Green taxi?
    - Are there any boroughs where the client is underperforming significantly in comparison to the competitors?
8. What are the peak hours for the taxi services for different taxi vendors?
    - Which vendor is most active in the evening?
9. Which taxi provider has the biggest proportion of disputed trips out of their total number of taxi trips?

4.4 Section D: Reporting in R Markdown

The aim of this section is to create a reproducible pdf report, containing the answers to the analysis from the Section C. This will be an R Markdown document which can be rendered into a pdf document.

The business requirements are provided below:

1. The R Markdown file should have coded in parameters which will allow a user to:
    - Input their own name and have it displayed in the report subtitle
    - Automatically display the current date each time the is generated, shown in the subtitle
    - Set the project path to the solution structure created in Section C
2. The generated pdf document should have the following structure:
    - A project title
    - A project subtitle stating who is the assessor and when the report was created
    - Table of contents
    - Clearly divided into sections, split by business question, with headers and sub headers
3. Use the path parameters to allow the user to source the code from the Section C with the solutions to the business questions. Any R code chunk you use for this should not be visible in the final report. Sourcing the external R file will prevent unnecessary copying of the existing code to generate the summary tables.
4. The final report should display both the code and the code's output, but no warnings or general messages caused by the program. The answers to the business questions and observations about the data should be clearly formulated in accompanying text chunks.
5. The theme and other stylistic choices are left to the programmer.

Save the RMD file in the PROGRAMS folder.

4.5 Section E: Hypothesis testing - *optional extension*

Carry out appropriate statistical test to answer the question number 2 from Section C. Recap of the question:

2. *Do any of the following features have an effect on the tip amount for Green Taxi? Hint: use a scatter plot to investigate the relationship between the trip distance and the tip amount.*
    - *Pickup hour*
    - *Payment type*
    - *Passenger count*
    - *Pick up borough*
    - *Trip distance*

The R syntax for these statistical tests is not covered by the R training but the theoretical knowledge of the hypothesis tests was covered by other courses. This section therefore requires additional research into what R code is needed for the selected statistical tests and which packages are needed.

Do not forget to formulate your null and alternative hypothesis and set the alpha levels before you carry out the relevant statistical tests. Any assumptions about the tests and the data should be included in the code's comments.

Create an R source file for this section with a new header and save it in the PROGRAM folder.

5 Appendix: data dictionary

This data dictionary describes taxi trip data.

5.1 Taxi_trips table

Field Name	Description
UniqueID	A unique ID key for every trip
VendorID	A numeric code indicating a taxi vendor. **1= Green taxi** **2 = Uber** **3 = Other**
Store_and_fwd_flag	This flag indicates whether the trip record 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**
RatecodeID	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**
Passenger_count	The number of passengers in the vehicle.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.

Mta_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Tip_amount	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Ehail_fee	$2.00 extra charge for Uber customers who tail taxis using its app.
Improvement_surcharge	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
Total_amount	The total amount charged to passengers. Does not include cash tips.
PaymentType	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**
Trip_type	A numeric code indicating trip zones. **1= Inner city** **2= Outer city** **3= Other**
Congestion_surcharge	Total amount collected in trip for NYS congestion surcharge.

5.2 Taxi_time_location table

Field name	Description
UniqueID	A unique ID key for every trip.
Taxi_trip	A flag showing whether the date and time is pick up and drop off information.
Time	The date and time when the meter was engaged and disengaged.
Location_details	A flag indicating whether a location was a pickup or a drop off information.
Location_codeID	A unique numeric code representing pick up and drop off locations.

5.3 Taxi_zone_lookup table

Field name	Description
LocationID	A unique numeric code representing pick up and drop off locations.
Borough	The names of New York City's primary administrative units
Zone	The names of New York City's administrative zones.
Service_zone	The names of New York City's service zones.

Unlock the power of your data

info@katalyzedata.com

katalyzedata.com