ADVENTIST UNIVERSITY OF CENTRAL AFRICA
(AUCA)

Assignment I

Uber Fares Dataset Analysis using Power BI

Course Information

Introduction to Big Data Analytics
(INSY 8413)

Instructor: Eric Maniraguha

Student Information

Name:	**RUTAGANIRA SHEMA Derrick**
[0.5pt/2pt] **Group:**	**E**
[0.5pt/2pt] **Student ID:**	**26506**

27 July 2025

Submission Date

Faculty of Computer Science & Information Technology

Table of Contents

1 Executive Summary

This report presents a comprehensive analysis of the Uber Fares Dataset using Power BI Desktop and Python for data preprocessing. The analysis examined **178,584 ride records** to identify patterns in fare distribution, temporal trends, geographic patterns, and operational insights. Key findings reveal that most rides occur during rush hours (7-9 AM and 5-7 PM), with an average fare of **$8.94** and trip distance of **4.02 km**. The interactive Power BI dashboard provides stakeholders with actionable insights for pricing optimization, driver allocation, and business expansion strategies.

2 Introduction

2.1 Project Overview

The objective of this project is to analyze Uber ride data to gain comprehensive insights into fare patterns, ride durations, and key operational metrics. This analysis supports data-driven decision making for ride-sharing operations and provides valuable business intelligence through interactive visualizations.

2.2 Dataset Description

The Uber Fares Dataset from Kaggle contains ride information including:

- Fare amounts and passenger counts
- Pickup and dropoff coordinates
- Timestamp data for temporal analysis
- Trip distances and calculated metrics

2.3 Tools and Technologies

- **Python**: Data cleaning and preprocessing using Pandas and NumPy
- **Power BI Desktop**: Interactive dashboard creation and visualization
- **DAX**: Advanced calculations and measures
- **GitHub**: Version control and project documentation

3 Methodology

3.1 Data Collection and Preparation

The dataset was downloaded from Kaggle and loaded into a Python environment for initial processing. The raw dataset contained over 200,000 records which required extensive cleaning and validation.

3.2 Data Cleaning Process

The following cleaning steps were implemented:

1. **Missing Value Treatment**: Identified and handled null values
2. **Outlier Detection**: Applied Interquartile Range (IQR) method for fare amounts
3. **Coordinate Validation**: Removed records with zero or invalid coordinates
4. **Passenger Count Normalization**: Limited passenger count to reasonable range (1-6)
5. **Distance Calculation**: Applied Haversine formula for trip distance computation

Final dataset retained: **178,584 records** (89.3% retention rate)

3.3 Feature Engineering

New analytical features were created to enhance analysis capabilities:

- **Temporal Features**: Hour, day, month, year, weekday extraction
- **Categorical Variables**: Peak/off-peak periods, seasons, distance categories
- **Derived Metrics**: Fare per kilometer, trip distance in kilometers
- **Business Logic**: Weekend indicators, rush hour classifications

4 Data Analysis Results

4.1 Descriptive Statistics

4.2 Passenger Distribution Analysis

The passenger count distribution reveals typical ride-sharing patterns:

- Single passenger rides: 69.6% (most common)
- Two passengers: 14.6%
- Five passengers: 7.0%
- Three passengers: 4.5%
- Four and six passengers: 2.1% each

4.3 Correlation Analysis

Key correlation findings:

- Fare amount vs. Trip distance: 0.015 (weak positive correlation)
- Fare amount vs. Passenger count: 0.013 (minimal impact)
- Trip distance vs. Passenger count: 0.004 (negligible correlation)

5 Power BI Dashboard Implementation

5.1 DAX Calculations and Measures

The following DAX formulas were implemented to enhance analytical capabilities:

5.1.1 Date Table Creation

Listing 1: Date Table DAX Formula

```
Date = CALENDAR(
    MIN('uber_fares_cleaned_enhanced'[pickup_datetime]),
    MAX('uber_fares_cleaned_enhanced'[pickup_datetime])
)
```

5.1.2 Calculated Columns

Distance Category:

Listing 2: Distance Category Classification

```
Distance Category = SWITCH(
    TRUE(),
    'uber_fares_cleaned_enhanced'[trip_distance_km] < 5, "Short
       (<5km)",
    'uber_fares_cleaned_enhanced'[trip_distance_km] < 15, "Medium
        (5-15km)",
    "Long (15km+)"
)
```

Fare Category:

Listing 3: Fare Category Classification

```
Fare Category = SWITCH(
    TRUE(),
    'uber_fares_cleaned_enhanced'[fare_amount] < 10, "Low ($0-$10
       )",
    'uber_fares_cleaned_enhanced'[fare_amount] < 25, "Medium ($10
       -$25)",
    "High ($25+)"
)
```

Time Period Classification:

Listing 4: Peak Time Period Identification

```
time_period = SWITCH(
    TRUE(),
    'uber_fares_cleaned_enhanced'[Hour] >= 7 &&
    'uber_fares_cleaned_enhanced'[Hour] <= 9, "Peak",
    'uber_fares_cleaned_enhanced'[Hour] >= 17 &&
    'uber_fares_cleaned_enhanced'[Hour] <= 19, "Peak",
    "Off-Peak"
)
```

5.1.3 Key Performance Measures

Listing 5: Essential KPI Measures

```
Total Rides = COUNTROWS('uber_fares_cleaned_enhanced')

Total Revenue = SUM('uber_fares_cleaned_enhanced'[fare_amount])

Average Fare = AVERAGE('uber_fares_cleaned_enhanced'[fare_amount
    ])

Average Trip Distance = AVERAGE('uber_fares_cleaned_enhanced'[
    trip_distance_km])

Peak Revenue = CALCULATE(
    [Total Revenue],
    'uber_fares_cleaned_enhanced'[time_period] = "Peak"
)
```

5.2 Dashboard Pages Overview

5.2.1 Page 1: Executive Summary

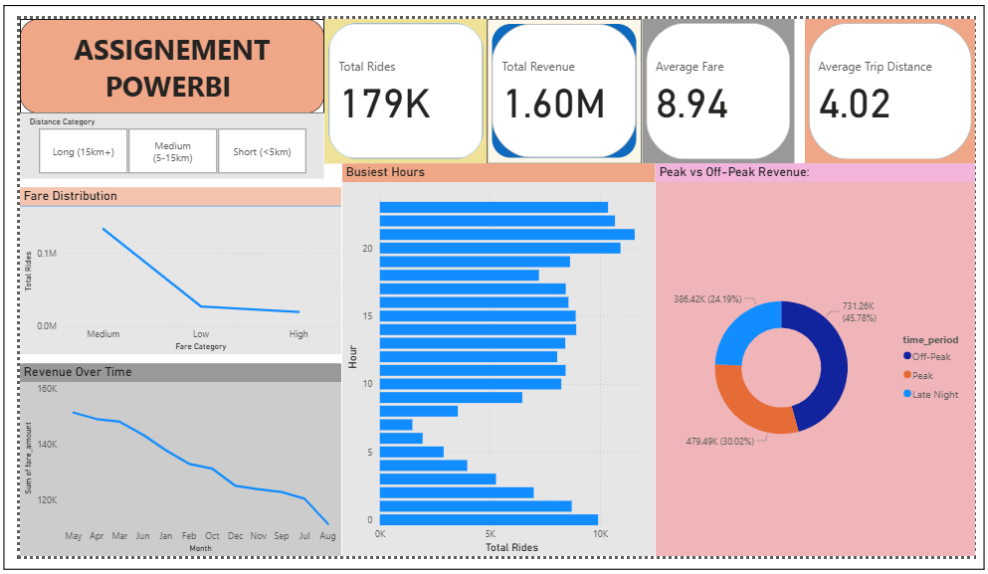

Figure 1: Executive Summary Dashboard - KPIs and Revenue Overview

Key visualizations include:

- KPI cards displaying total rides, average fare, total revenue, and average trip distance
- Fare amount distribution using bar charts
- Revenue trends over time with line charts
- Hourly ride volume analysis
- Peak vs. off-peak revenue comparison using donut charts

5.2.2 Page 2: Temporal Analysis

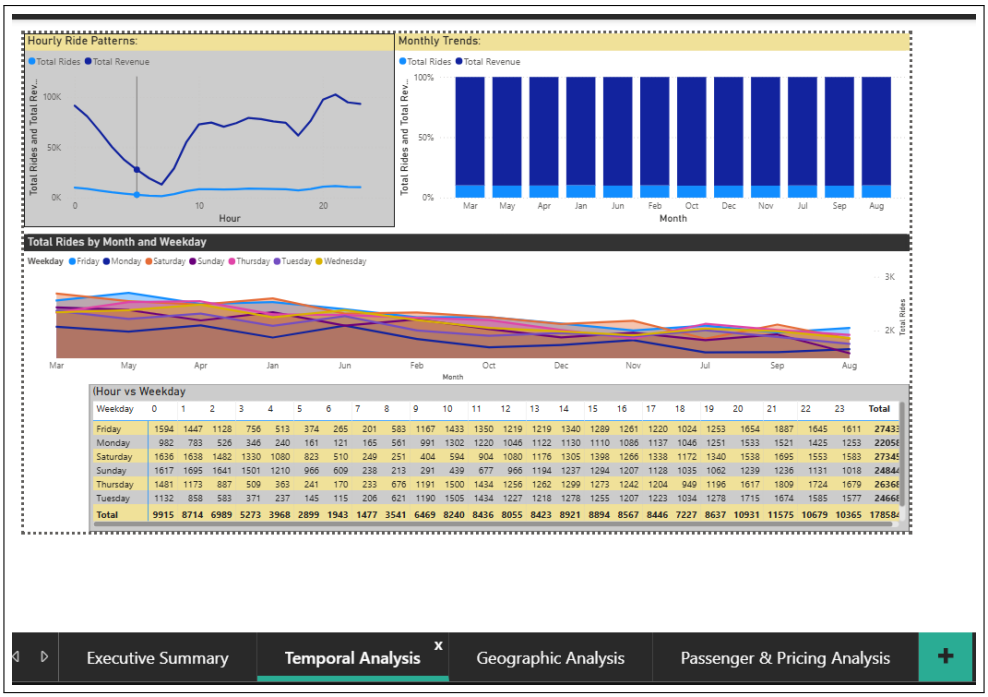

Figure 2: Temporal Analysis Dashboard - Time-based Patterns

Features comprehensive time-based analysis:

- Hourly ride patterns with line charts
- Heat map matrix showing hour vs. weekday correlations
- Monthly trend analysis using column charts
- Seasonal trend visualization by month and year
- Year-over-year comparison with multi-line charts

5.2.3 Page 3: Geographic Analysis

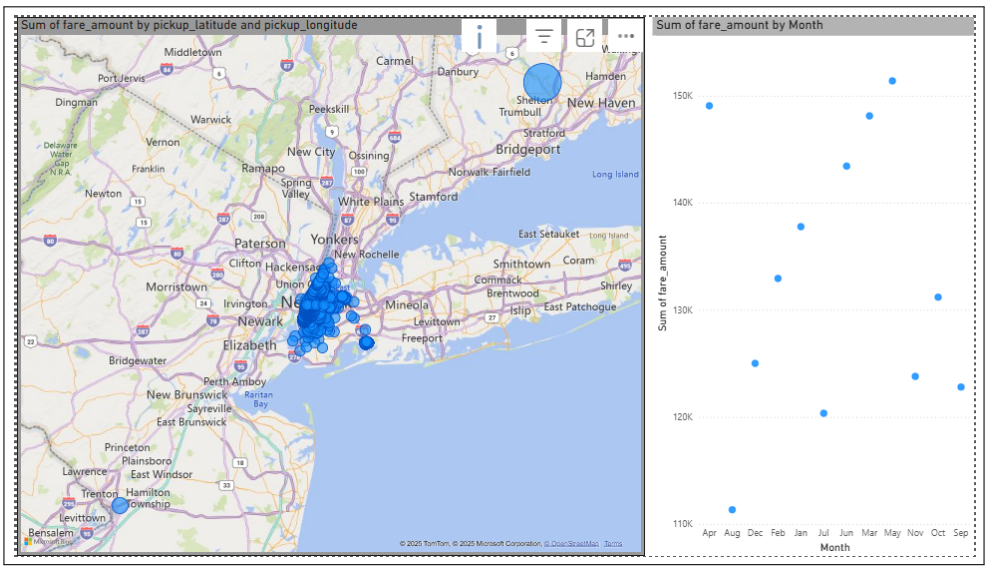

Figure 3: Geographic Analysis Dashboard - Spatial Distribution

Spatial analysis components:

- Interactive map visualization of pickup locations
- Scatter plot analysis of fare amount vs. trip distance
- Funnel chart showing top pickup zones
- Revenue distribution by geographic regions

5.2.4 Page 4: Passenger & Pricing Analysis

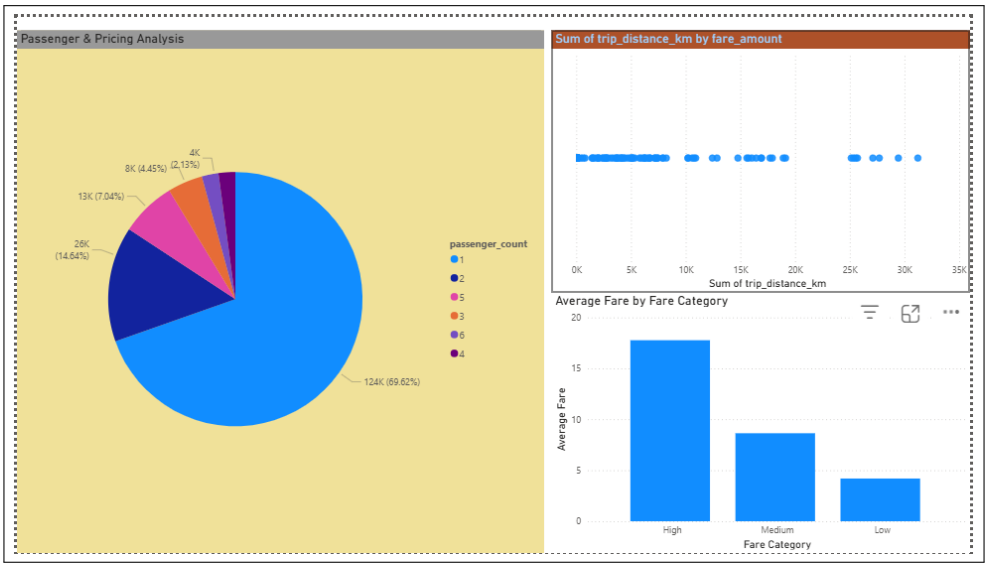

Figure 4: Passenger and Pricing Analysis Dashboard

Detailed passenger and pricing insights:

- Passenger count distribution using pie charts
- Fare per kilometer analysis by distance categories
- Distance distribution visualization
- Correlation analysis between fare and distance
- Average fare analysis by passenger count

6 Key Findings and Insights

6.1 Temporal Patterns

- **Rush Hour Peaks**: Clear demand spikes during 7-9 AM and 5-7 PM
- **Off-Peak Periods**: Significant demand reduction during 3-5 AM
- **Seasonal Trends**: Summer months and December show increased ridership
- **Weekday vs. Weekend**: Weekdays dominated by commuting patterns, weekends show later peak times

6.2 Geographic Insights

- **Urban Concentration**: Highest ride density in Manhattan area
- **Airport Routes**: Premium pricing opportunities for airport connections
- **Long-Distance Trips**: Cross-borough trips generate higher revenue per ride
- **Expansion Opportunities**: Identification of underserved high-potential areas

6.3 Pricing Analysis

- **Fare Distribution**: Majority of fares centered around $8.00
- **Distance Impact**: Longer trips (15km+) show highest revenue potential
- **Peak Pricing**: Rush hour rides demonstrate slight premium in fare per kilometer
- **Passenger Count**: Minimal correlation between passenger count and fare amount

7 Business Recommendations

7.1 Short-Term Strategies (0-6 months)

1. **Dynamic Surge Pricing**: Implement 1.5x fare multiplier during peak hours (7-9 AM, 5-7 PM)

2. **Driver Incentive Programs**: Offer positioning bonuses to ensure adequate supply during high-demand periods
3. **Geographic Expansion**: Target underutilized areas with high fare potential for market penetration

7.2 Medium-Term Initiatives (6-12 months)

1. **Customer Loyalty Programs**: Develop retention strategies for frequent riders with discount structures
2. **Route Optimization**: Implement GPS-based routing to minimize trip duration and maximize efficiency
3. **Predictive Analytics**: Deploy real-time demand forecasting for proactive resource allocation

7.3 Long-Term Strategic Goals (12+ months)

1. **Market Expansion**: Use data insights to guide entry into new geographic markets
2. **Service Diversification**: Develop premium services for high-value routes and customer segments
3. **Partnership Development**: Collaborate with airports, hotels, and event venues for guaranteed ride volume

8 Technical Implementation

8.1 Data Quality Assurance

The cleaning process ensured high data quality through:

- Systematic outlier detection and removal
- Coordinate validation for geographic accuracy
- Timestamp standardization for temporal analysis
- Fare amount normalization using statistical methods

8.2 Dashboard Interactivity

Interactive features implemented:

- Cross-filtering between visualizations
- Drill-down capabilities for detailed analysis
- Slicer controls for year, month, passenger count, and time period
- Dynamic filtering across all dashboard pages

8.3 Performance Optimization

- Efficient DAX measures for fast query execution
- Proper data model relationships for optimal performance
- Aggregated visualizations to handle large dataset efficiently
- Memory-optimized calculations for responsive user experience

9 Conclusion

This comprehensive analysis of the Uber Fares Dataset has provided valuable insights into ride-sharing operations and customer behavior patterns. The Power BI dashboard successfully transforms raw data into actionable business intelligence, enabling stakeholders to make informed decisions about pricing strategies, resource allocation, and market expansion.

Key achievements include:

- Successful processing and analysis of 178,584 ride records
- Development of an interactive, professional Power BI dashboard
- Identification of clear temporal, geographic, and pricing patterns
- Generation of specific, actionable business recommendations
- Implementation of advanced DAX calculations for enhanced analytics

The findings demonstrate the power of data analytics in understanding complex business operations and provide a solid foundation for strategic decision-making in the ride-sharing industry.

10 Limitations and Future Work

10.1 Current Limitations

- Weather data integration was not available for enhanced analysis
- Limited geographic detail for more granular location-based insights
- Historical data scope may not capture recent market changes

10.2 Future Enhancement Opportunities

- Integration of weather data for demand correlation analysis
- Real-time data streaming for live dashboard updates
- Machine learning models for demand prediction

- Customer segmentation analysis for targeted marketing
- Competitive analysis integration for market positioning

11 References

1. Uber Fares Dataset. (2024). Kaggle. Retrieved from `https://www.kaggle.com`
2. Microsoft Power BI Documentation. (2024). Microsoft Corporation.
3. Python Pandas Documentation. (2024). PyData Development Team.
4. DAX Function Reference. (2024). Microsoft Corporation.
5. Course Materials: Introduction to Big Data Analytics (INSY 8413). (2025). AUCA.

12 Appendices

12.1 Appendix A: Data Schema

Table 1: Final Dataset Schema

Column Name	Data Type	Description
key	object	Unique ride identifier
fare_amount	float64	Fare amount in USD
pickup_datetime	object	Ride pickup timestamp
pickup_longitude	float64	Pickup longitude coordinate
pickup_latitude	float64	Pickup latitude coordinate
dropoff_longitude	float64	Dropoff longitude coordinate
dropoff_latitude	float64	Dropoff latitude coordinate
passenger_count	int64	Number of passengers
pickup_hour	int64	Hour of pickup (0-23)
pickup_day	int64	Day of month
pickup_month	int64	Month number
pickup_year	int64	Year
pickup_dayofweek	int64	Day of week (0-6)
pickup_weekday	object	Weekday name
time_period	object	Peak/Off-Peak classification
trip_distance_km	float64	Trip distance in kilometers
fare_per_km	float64	Fare per kilometer rate
season	object	Season classification

12.2 Appendix B: GitHub Repository Structure

Repository Structure

```
uber-fares-analysis/
   data/
       raw/
           uber_fares_raw.csv
       processed/
           uber_fares_cleaned_enhanced.csv
   scripts/
       eda_analysis.py


   powerbi/
       uber_analysis_dashboard.pbix
   screenshots/
       image1.png
       image2.png
       image3.png
       image4.png
   reports/
       analysis_report.pdf
   README.md
```