

**IBM SPSS Modeler**

Project Report

On

**“Analysis of Uber’s Ride Data”**

**Submitted by:** Anurag Mahato

Priyanka Chandra

**Course:** B.Tech CSE (AI & DS)

**Semester:** 5th

**Institution:** Arka Jain University

**Guide:** Mr. Prajwal Deep Khokhar

**Date:** 25-10-2025

**Index**

|  |  |  |
| --- | --- | --- |
| **S.no** | **Topic** | **Pg. no** |
| 1. | Project Brief | 3 |
| 2. | Introduction | 4 |
| 3. | Feasibility Study | 5 |
| 4. | Project Details | 6 |
|  | 4.1 About the Dataset | 6 |
|  | 4.2 Workflow | 7 |
| 5. | Conclusion / Summary | 10 |

1. **Project Brief**

This project explores how data analytics can be used to understand and improve the operations of app-based cab services.

Using **IBM SPSS Modeler**, three analytical models were developed :

* **Logistic Regression Model** for predicting the *Booking Status*, and
* **K-Means** and **Kohonen Models** for *Customer Segmentation*.

The **Logistic Regression Model** focuses on predicting whether a ride will be *completed* or *cancelled* due to different reasons, based on influencing factors like **vehicle type, waiting time, and customer ratings**. This helps identify key causes behind booking failures and allows businesses to take corrective actions.

In addition, **K-Means** and **Kohonen clustering models** group customers with similar ride patterns — such as **ride distance, booking value, and trip duration** — into distinct clusters. These clusters provide useful insights for **targeted marketing, service personalization**, and **strategic planning**.

Together, these models showcase how predictive and clustering techniques can transform raw ride data into meaningful insights that drive **efficiency, customer satisfaction, and business growth** in the ride-hailing industry.

1. **Introduction**

With the rise of app-based cab services like Uber, analyzing customer behavior and predicting booking outcomes has become essential for improving service quality and operational efficiency. Each ride generates large volumes of valuable data that can reveal important business insights.

The goal is to uncover insights that can enhance decision-making, customer satisfaction, and overall business performance.

Using SPSS Modeler, the project demonstrates how data can be transformed into actionable knowledge through:

* **Predictive Model (Logistic Regression),** that predicts the *Booking Status* of rides — whether completed or cancelled (and for what reason).
  + This helps Uber identify cancellations and the key factors influencing them — such as long waiting times, driver unavailability, or specific vehicle types — allowing the company to take preventive actions and improve service reliability.
* The **Clustering Models** (**K-Means** and **Kohonen)**, that perform *customer segmentation* by grouping customers based on **ride distance, booking value, and vehicle type**.
  + These clusters help Uber design better offers, personalize services, and optimize pricing strategies.

The entire process follows the **CRISP-DM** methodology, covering all stages — from **data understanding and preparation** to **modeling, evaluation, and insight generation**. Data quality was maintained using **Data Audit, Derive, Reclassify, and Type nodes** before building the models.

1. **Feasibility study**

* **Technical Feasibility**  
  The project is technically feasible as it was developed using **IBM SPSS Modeler**, a robust analytical tool that supports a visual, drag-and-drop interface. This eliminates the need for complex programming.
* **Operational Feasibility**The workflow designed in SPSS Modeler is simple, logical, and easy to execute. Each step — from **data cleaning and transformation** to **model building and evaluation** — was carried out through well-connected nodes. The models are interpretable and can be directly applied for **business decision-making**, such as identifying reasons for cancellations or understanding customer clusters.
* **Data** **Feasibility**

The dataset provided adequate records and relevant attributes, which were sufficient to build meaningful predictive and clustering models. The predictive model achieved an accuracy of 96.59%, and the two clustering models maintained a silhouette score of 0.5, validating the reliability and practicality of the results.

Overall, the project demonstrates a feasible and scalable framework for predictive analysis and customer segmentation in ride-hailing services.

1. **Project Details**
   1. **About the Dataset**

**Uber Ride Analytics Dataset (2024):**

* File: ncr\_ride\_bookings.csv
* Rows: 1,50,000
* Columns: 21 fields covering booking, customer, and ride details
* Description:  
  The dataset contains detailed information on Uber ride bookings across various vehicle types and locations. It includes both successful and cancelled trips, offering insights into booking behavior and ride outcomes.

Key attributes include:

* + Ride Distance
  + Average Driver Arrival Time (Avg VTAT)
  + Trip Duration
  + Booking Value
  + Vehicle Type
  + Driver and Customer Ratings
  + Booking Status — with categories such as Completed, Cancelled by Customer, Cancelled by Driver, Driver Not Found, and Incomplete.

Fields Used for Modeling:

* Predictive Model (Logistic Regression):  
  The target field used was Booking Status, allowing the model to predict whether a ride would be completed or cancelled (and for what reason).
* Clustering Models (K-Means & Kohonen):  
  Key numerical fields such as Ride Distance, Booking Value, Vehicle Type and Avg CTAT (converted from categorical) were used to group customers with similar ride patterns and preferences.
  1. **Workflow**

**🔹 Logistic Regression Model:**

**Objective:** To predict Booking Status (cancelled or completed, and for what reason).

**Data Flow:**

[ Data Understanding ]

* 1. **Var File Node** – Imported the Uber dataset.
  2. **Data Audit Node** – Checked data quality, measurement variables and identified missing or invalid values.

[ Data Preparation ]

* 1. **Derive Node** – Replaced null values with 0 and stored the data into new fields of numeric type.
  2. **Reclassify Node** (optional) – Created a new field Booking\_Status\_(f*lag)* where 1 represents completed rides and 0 represents cancelled rides.
  3. **Filter Node** – Filtered out unnecessary fields.
  4. **Type Node** – Defined roles of variables (target and inputs).

[ Modeling ]

* 1. **Partition Node** – Divided the dataset into 60-40 (60% for training and 40% for testing).
  2. **Logistic Regression Node** – Built the model to predict booking success probability.

[ Evaluation ]

* 1. **Analysis Node** – Interpreted the model accuracy and variable importance.

**Interpretation**:

The output fields like **$L-Booking Status** and **$LP-Booking Status** represent predicted labels and probabilities respectively. The model showed that factors such as **waiting time (Avg\_VTAT)**, **ratings** and **booking value** had significant influence on the cancellation likelihood.

**🔹 K-Means Model:**

**Objective:** To segment customers into groups based on ride and booking patterns.

**Data Flow:**

[ Data Understanding ]

(Same as that of the Logistic Regression Model)

[ Data Preparation ]

* 1. **Derive Node** – Replaced null values with 0 and stored the data into new fields of numeric type.
  2. **Filter Node** – Selected relevant fields for clustering.
  3. **Reclassify Node** – Assigned numeric codes (1–7) to different vehicle types, creating a new field *Vehicle\_Type\_Code*.
  4. **Type Node** – Set appropriate variable roles (input/target).

[ Modeling ]

* 1. **K-Means Node** – Built the clustering model with a **silhouette score of 0.5**, indicating good cluster separation.

[ Evaluation ]

* 1. **3D Plot Node** – Visualized results with:
     + **X-axis:** Vehicle\_Type\_Code
     + **Y-axis:** Ride Distance(num)
     + **Z-axis:** Booking Value(num)
     + **Overlay color:** $KM-Kmeans\_customer\_seg

**Interpretation:**  
The K-Means model grouped customers based on their **spending**, **ride distance**, and **vehicle preferences**. It helps in identifying customer segments like high-value riders, short-trip customers, and long-distance travelers, which can support targeted marketing and service optimization.

**🔹 Kohonen Model:**

**Objective:** To perform neural-based clustering for discovering deeper patterns in customer data.

**Data Flow:**

[ Data Understanding ]

(Same as that of the Logistic Regression Model)

[ Data Preparation ]

* 1. **Derive Node** – Replaced null values with 0 and stored the data into new fields of numeric type.
  2. **Filter Node** – Selected relevant fields for clustering.
  3. **Type Node** – Defined variable roles (input/target).

[ Modeling ]

* 1. **Kohonen Node** – Built the model with a **silhouette score of 0.5**, similar to K-Means, ensuring optimal cluster formation.

[ Evaluation ]

* 1. **2D Plot Node** – Represented results with:
     + **X-axis:** Booking Value(num)
     + **Y-axis:** Ride Distance(num)
     + **Overlay color:** $KXY-Kohonen\_customer\_seg

**Interpretation:**  
The Kohonen model (Self-Organizing Map) provided a two-dimensional representation of high-dimensional customer data. It grouped similar customers close to each other in the plot, offering insights that complement the K-Means results.

1. **Conclusion/Summary**

This project successfully implemented three analytical models — Logistic Regression, K-Means, and Kohonen — to address different objectives within the Uber dataset.

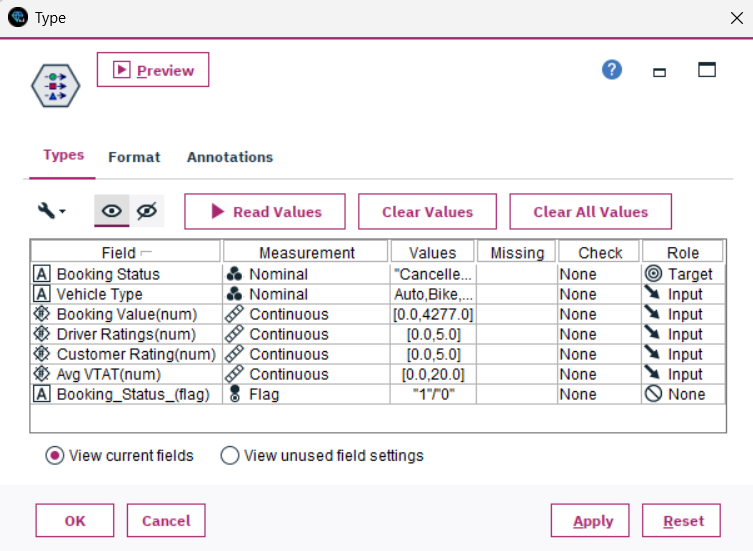
* The **Logistic Regression model** accurately predicted booking outcomes, revealing that longer waiting times, vehicle type, and ratings significantly influenced cancellations.

Figure 1: *Type node showing the defined data roles for the Logistic model.*

Figure 2: *Analysis node results of Logistic Regression model.*

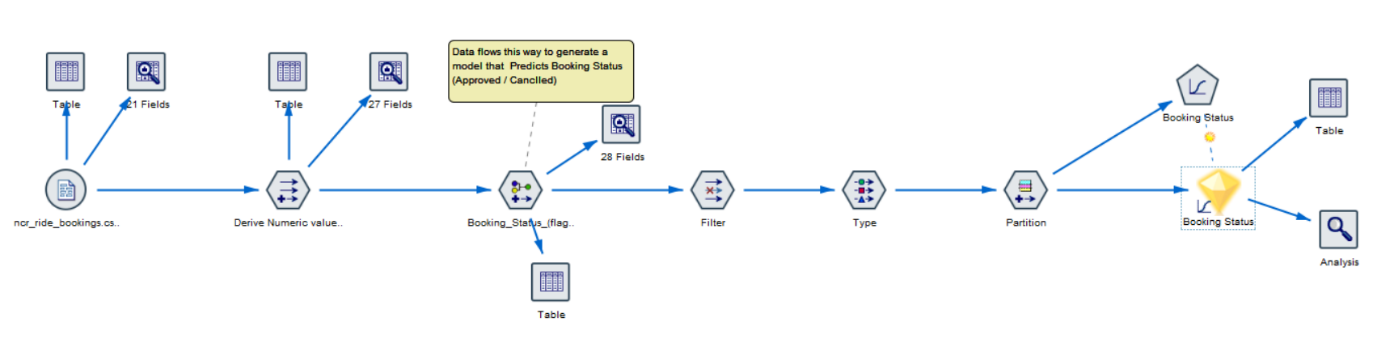


Figure 3: *Workflow of Logistic Regression model.*

* The **K-Means and Kohonen models** effectively segmented customers into meaningful clusters, offering insights for personalized marketing and service enhancement. They grouped users with similar trip distances, booking values and vehicle types, helping the company to design better offers and improve customer service.

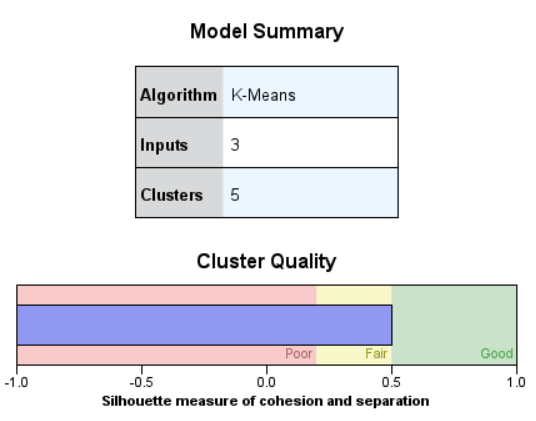


Figure 4: *Model Summery and Cluster Quality of K-Means model.*

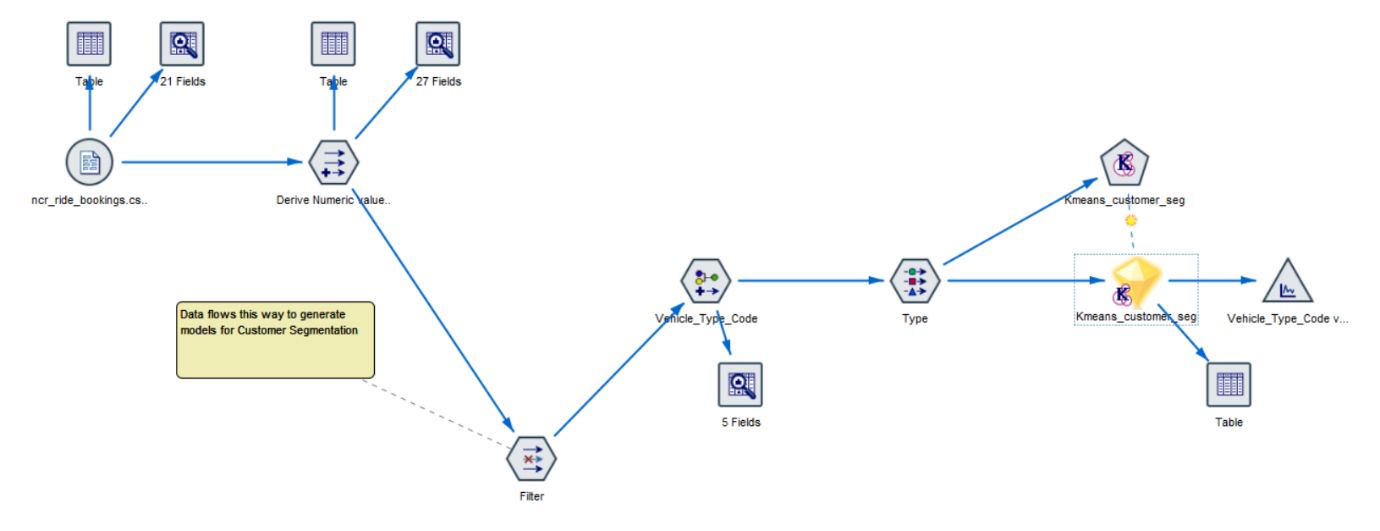


Figure 5: *Workflow of K-Means model.*

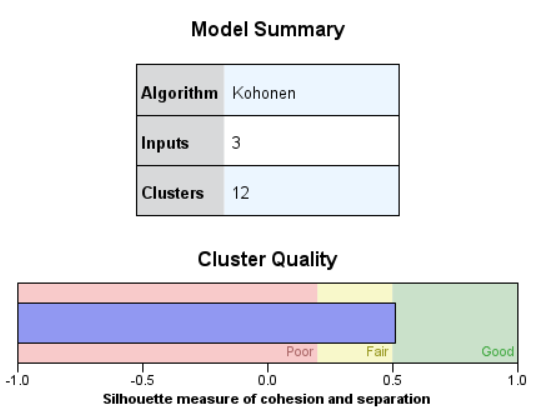


Figure 6: *Model Summery and Cluster Quality of Kohonen model.*

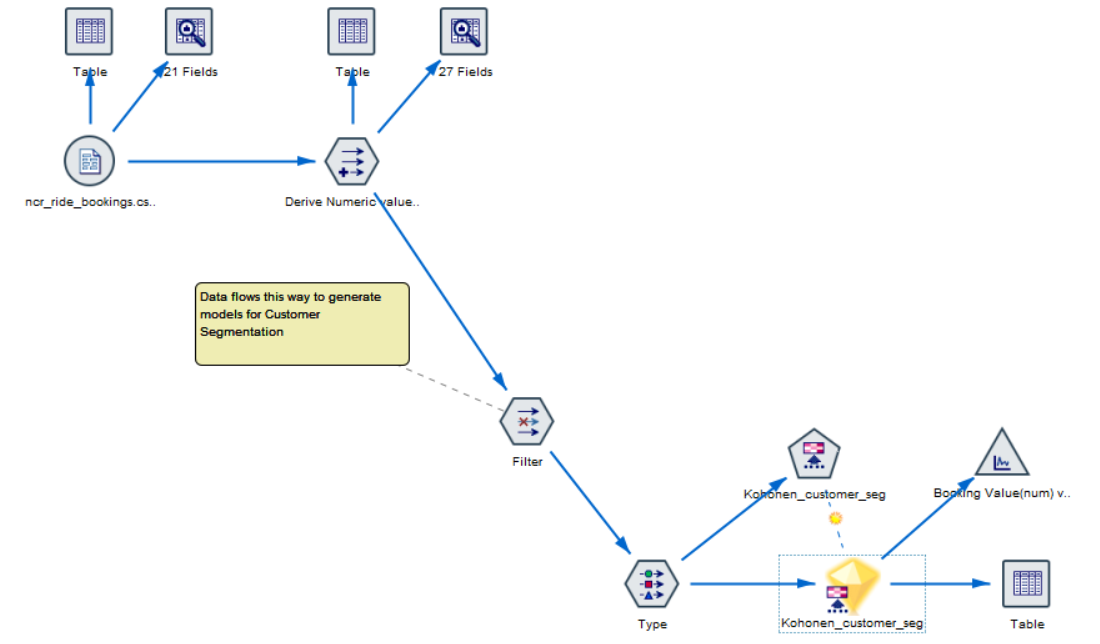
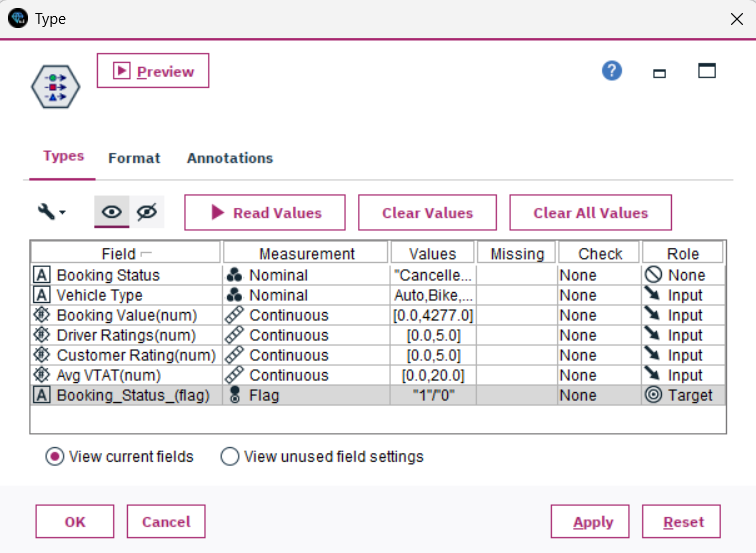


Figure 7: *Workflow of Kohonen model.*

By combining predictive and clustering approaches, the project highlights the power of data analytics in understanding customer behavior, improving business operations, and reducing service inefficiencies.

During the development of Logistic Regression model, it was noticed that,

* when we use the **reclassified** flag field **Booking\_Status\_(flag)** as the target, it gives the result with **100% accuracy**.
* However, this accuracy was **not meaningful**.  
  By simply classifying rides as “**Completed**” or “**Cancelled**,” the model could not show the *reason* behind cancellations.
* For a company, knowing *why* a booking was cancelled — for example, “no driver found,” “cancelled by driver,” or “cancelled by customer” — is far more useful than just knowing that it was cancelled.
* That’s why the **Booking Status field** was chosen as the target instead.  
  Even though this slightly reduced the accuracy, it made the model **more valuable and practical** for real-world decision-making.

  
  
 Figure 8: *Type node* ***in case we used Booking\_Status\_(flag) as Target****.*

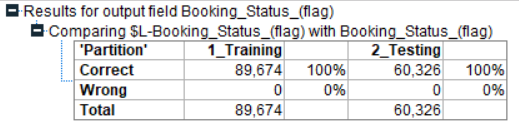


Figure 9: *Analysis results* ***in case we used Booking\_Status\_(flag) as Target****.*