A PROJECT REPORT

on

“Amazon Logistics Delivery Performance Analysis”

Using Excel

Table of Contents

Abstract

The project titled **"Amazon Logistics Delivery Performance Analysis"** was carried out to apply data analytics concepts to a real-world logistics dataset and derive meaningful insights about delivery efficiency. The dataset, sourced from Kaggle, contained information about Amazon delivery operations including order details, agent demographics, vehicle types, weather and traffic conditions, delivery times and geographical coordinates.

The analysis was conducted entirely in **Microsoft Excel**, following a structured approach that included **data cleaning**, **pre-processing**, **feature engineering**, **descriptive statistics**, **pivot-based exploration** and interactive **dashboard development**. The objective was to identify the key factors that influence delivery performance and understand patterns in delivery time, waiting time and operational efficiency.

Through this analysis, several critical insights were discovered. It was observed that **heavy traffic** and **unfavourable weather conditions** significantly increased delivery time. Deliveries conducted during **early morning** and **late-night hours** or in **urban regions** showed higher waiting times, while younger agents (**<25 years of age**) maintained better average ratings and performance consistency. Furthermore, **distance** and **delivery time** were found to be positively correlated, confirming expected real-world behaviour.

The final outcome of the project was an **interactive dashboard** summarizing all key metrics such as average delivery time, waiting time, distance and agent rating. The dashboard allowed filtering through slicers for traffic, weather, vehicle type and delivery area enabling a dynamic exploration of operational patterns.

Overall, this project provided a comprehensive understanding of how **structured data handling and visualization** can transform raw information into actionable insights. It bridged the gap between theoretical learning and real-world data analysis, demonstrating how analytics supports data-driven decision-making in logistics.

1. Introduction

In the era of digital transformation, **data analytics** has become an essential tool for improving business efficiency, decision-making and customer satisfaction. Within the e-commerce industry, logistics forms the backbone of operations, where timely and accurate delivery is critical to maintaining service quality. Companies such as **Amazon** leverage data-driven insights to optimize routes, monitor agent performance and adapt to environmental challenges like traffic or weather fluctuations.

This project titled **"Amazon Logistics Delivery Performance Analysis"** was undertaken to replicate a real-world analytical workflow using an open-source logistics dataset. The dataset sourced from **Kaggle**, provided detailed records of delivery operations, including information such as order identifiers, delivery agent details, timestamps, vehicle types, environmental conditions and delivery durations.

The aim of the project was to apply analytical techniques in **Microsoft Excel** to identify the factors influencing delivery performance. By examining variables such as **distance**, **agent demographics**, **area type, traffic density** and **weather conditions**, the project sought to uncover patterns that affect delivery efficiency and waiting time.

The project was executed through a structured analytical workflow that included **data pre-processing**, **feature engineering**, **descriptive statistics**, **pivot-based exploration** and **interactive dashboard creation**. Each stage contributed to building a deeper understanding of how various operational and environmental factors interact within a delivery system.

The insights obtained from this project reflect the real-world challenges faced by logistics networks and illustrate how data analysis can be used to make such systems more efficient and reliable.

1.1 Objectives and Business Questions

Objective

The primary objective of this project was to analyze the Amazon Logistics dataset to identify operational trends, evaluate agent performance and determine the major factors influencing delivery efficiency, waiting time and overall service quality.

Business Questions

To achieve this objective, a set of analytical questions were framed to guide the analysis and align each stage of the workflow with a measurable outcome:

1. How does **delivery time** vary under different **traffic** and **weather conditions**?
2. Do certain **geographic areas** or **vehicle types** experience higher delivery durations?
3. Is there a correlation between the **agent's age** or **rating** and their delivery performance?
4. Which **time of day** (morning, afternoon, evening or night) records the highest waiting times or order volumes?
5. How does **delivery distance** affect total delivery and pickup times?
6. Which **product categories** (e.g. electronics, groceries) dominate the delivery count and do they influence delivery time?
7. What is the **average waiting time, delivery distance** and **agent rating** and how do they together define operational efficiency?
8. What patterns or insights can be drawn to recommend improvements in delivery management and agent allocation?

These business questions served as the foundation of the analysis, ensuring that each step from cleaning to visualization contributed toward addressing a clear operational goal. The project thus combined academic learning with practical application, providing a holistic view of how **data analytics can optimize logistics performance** in a real-world business environment.

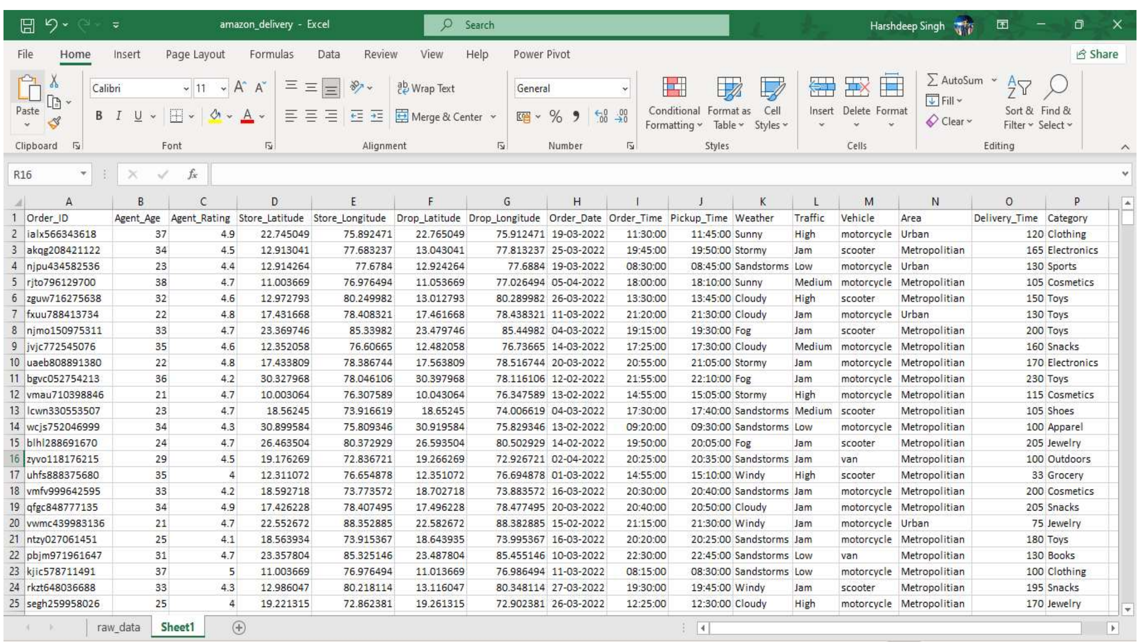
2. Dataset Description

The dataset used for this project was sourced from **Kaggle**, an open-source platform that hosts real-world datasets for data analysis and machine learning practice. The chosen dataset, titled **Amazon Delivery Logistics Dataset**, provided a detailed snapshot of Amazon's delivery operations, containing valuable information about orders, delivery agents, geographical details and environmental factors influencing logistics performance.

This dataset was ideal for analytical exploration because it contained both **quantitative** and **categorical** variables, enabling a comprehensive study of patterns, trends and correlations.

2.1 Overview of the Dataset

- **Source:** Kaggle **- Amazon Delivery Logistics Dataset**
- **File Format:** CSV (Comma-Separated Values)
- **Number of Records:** 43,739 rows (before cleaning and validation)
- **Number of Columns:** 15 (original) + 6 (engineered)
- **Tool Used for Analysis:** Microsoft Excel
- **Scope:** To analyze factors affecting delivery performance and identify optimization opportunities.



Order_ID	Agent_Age	Agent_Rating	Store_Latitude	Store_Longitude	Drop_Latitude	Drop_Longitude	Order_Date	Order_Time	Pickup_Time	Weather	Traffic	Vehicle	Area	Delivery_Time	Category
ialx566343618	37	4.9	22.745049	75.892471	22.765049	75.912471	19-03-2022	11:30:00	11:45:00	Sunny	High	motorcycle	Urban	120	Clothing
akqg208421122	34	4.5	12.913041	77.683237	13.043041	77.813237	25-03-2022	19:45:00	19:50:00	Stormy	Jam	scooter	Metropolitian	165	Electronics
njpu434582536	23	4.4	12.914264	77.6784	12.924264	77.6884	19-03-2022	08:30:00	08:45:00	Sandstorms	Low	motorcycle	Urban	130	Sports
rjto796129700	38	4.7	11.003669	76.976494	11.053669	77.026494	05-04-2022	18:00:00	18:10:00	Sunny	Medium	motorcycle	Metropolitian	105	Cosmetics
zguw716275638	32	4.6	12.972793	80.249982	13.012793	80.289982	26-03-2022	13:30:00	13:45:00	Cloudy	High	scooter	Metropolitian	150	Toys
fxuu788413734	22	4.8	17.431668	78.408321	17.461668	78.438321	11-03-2022	21:20:00	21:30:00	Cloudy	Jam	motorcycle	Urban	130	Toys
njmo150975311	33	4.7	23.369746	85.33982	23.479746	85.44982	04-03-2022	19:15:00	19:30:00	Fog	Jam	scooter	Metropolitian	200	Toys
jvjc772545076	35	4.6	12.352058	76.60665	12.482058	76.73665	14-03-2022	17:25:00	17:30:00	Cloudy	Medium	motorcycle	Metropolitian	160	Snacks
uaeb808891380	22	4.8	17.433809	78.386744	17.563809	78.516744	20-03-2022	20:55:00	21:05:00	Stormy	Jam	motorcycle	Metropolitian	170	Electronics
bgvc052754213	36	4.2	30.327968	78.046106	30.397968	78.116106	12-02-2022	21:55:00	22:10:00	Fog	Jam	motorcycle	Metropolitian	230	Toys
vmau710398846	21	4.7	10.003064	76.307589	10.043064	76.347589	13-02-2022	14:55:00	15:05:00	Stormy	High	motorcycle	Metropolitian	115	Cosmetics
lcwn330553507	23	4.7	18.56245	73.916619	18.65245	74.006619	04-03-2022	17:30:00	17:40:00	Sandstorms	Medium	scooter	Metropolitian	105	Shoes
wcjs752046999	34	4.3	30.899584	75.809346	30.919584	75.829346	13-02-2022	09:20:00	09:30:00	Sandstorms	Low	motorcycle	Metropolitian	100	Apparel
blhl288691670	24	4.7	26.463504	80.372929	26.593504	80.502929	14-02-2022	19:50:00	20:05:00	Fog	Jam	scooter	Metropolitian	205	Jewelry
zyvo118176215	29	4.5	19.176269	72.836721	19.266269	72.926721	02-04-2022	20:25:00	20:35:00	Sandstorms	Jam	van	Metropolitian	100	Outdoors
uhfs888375680	35	4	12.311072	76.654878	12.351072	76.694878	01-03-2022	14:55:00	15:10:00	Windy	High	scooter	Metropolitian	33	Grocery
vmfv999642595	33	4.2	18.592718	73.773572	18.702718	73.883572	16-03-2022	20:30:00	20:40:00	Sandstorms	Jam	motorcycle	Metropolitian	200	Cosmetics
qfgc848777135	34	4.9	17.426228	78.407495	17.496228	78.477495	20-03-2022	20:40:00	20:50:00	Cloudy	Jam	motorcycle	Metropolitian	205	Snacks
vwmc439983136	21	4.7	22.552672	88.352885	22.582672	88.382885	15-02-2022	21:15:00	21:30:00	Windy	Jam	motorcycle	Urban	75	Jewelry
ntzy027061451	25	4.1	18.563934	73.915367	18.643935	73.995367	16-03-2022	20:20:00	20:25:00	Sandstorms	Jam	motorcycle	Metropolitian	180	Toys
pbjm971961647	31	4.7	23.357804	85.325146	23.487804	85.455146	10-03-2022	22:30:00	22:45:00	Sandstorms	Low	van	Metropolitian	130	Books
kjic578711491	37	5	11.003669	76.976494	11.013669	76.986494	11-03-2022	08:15:00	08:30:00	Sandstorms	Low	motorcycle	Metropolitian	100	Clothing
rkzt648036688	33	4.3	12.986047	80.218114	13.116047	80.348114	27-03-2022	19:30:00	19:45:00	Windy	Jam	scooter	Metropolitian	195	Snacks
segh259958026	25	4	19.221315	72.862381	19.261315	72.902381	26-03-2022	12:25:00	12:30:00	Cloudy	High	motorcycle	Metropolitian	170	Jewelry

***Figure 2.1.1**: Dataset Overview*

The dataset contained a rich combination of numerical, categorical, spatial and temporal data fields. Such diversity made it possible to analyze the interplay between internal variables (like agent characteristics) and external variables (like traffic and weather).

2.2 Original Dataset Columns

The table below lists and explains the key variables present in the dataset before any modifications or transformations were performed.

Column Name	Data Type	Description
Order_ID	Text	Unique identifier for each delivery order.
Agent_Age	Numeric	Age of the delivery agent in years.
Agent_Rating	Numeric	Rating of the delivery agent given by customers, typically on a 1–5 scale.
Store_Latitude	Numeric	Latitude coordinate of the store or pickup location.
Store_Longitude	Numeric	Longitude coordinate of the store or pickup location.
Drop_Latitude	Numeric	Latitude coordinate of the delivery drop point.
Drop_Longitude	Numeric	Longitude coordinate of the delivery drop point.
Order_Date	Date	The date when the order was placed.
Order_Time	Time	The time at which the order was received by the system.
Pickup_Time	Time	The time when the agent picked up the order for delivery.
Weather	Categorical	Describes the weather condition during delivery (e.g., Sunny, Fog, Cloudy etc).
Traffic	Categorical	Indicates traffic density during delivery (Low, Medium, High, Jam).
Vehicle	Categorical	Type of vehicle used for delivery (e.g., Motorcycle, Scooter, Car)
Area	Categorical	Geographical area classification (Urban, Semi-Urban, or Metropolitan).
Delivery_Time	Numeric	Total delivery time in minutes (dependent variable for performance).
Category	Categorical	The product category of the item delivered (e.g., Books, Electronics, Grocery etc.).

2.3 Derived Columns (Feature-Engineered Attributes)

During the course of analysis, several new variables were created to extract deeper insights and simplify comparisons. These engineered features are summarized below:

Derived Column	**Purpose**
Day_Name	Extracted the weekday name from the order date to analyze daily order trends.
Wait_Time_Min	Calculated the waiting time (in minutes) between order placement and pickup.
Order_Hour	Extracted the hour from order time to analyze hourly delivery patterns.
Hour_Bucket	Grouped hours into time categories (Morning, Afternoon, Evening, Night).
Agent_Age_Bucket	Categorized agents into age ranges to compare performance across demographics.
Distance_Km	Calculated straight-line distance (in kilometers) between store and delivery location using the Haversine formula.

These engineered columns enhanced the dataset's analytical depth, making it easier to uncover patterns across time, geography and performance attributes.

2.4 Data Quality and Suitability

The dataset was examined for accuracy, completeness and consistency. Missing values were treated appropriately, duplicates were removed and invalid time entries were corrected. The data was found to be sufficiently large and diverse to support meaningful statistical analysis and visualization.

The combination of agent-specific, temporal, spatial and environmental data provided a **360-degree view of delivery operations**, enabling in-depth exploration of how various conditions affect delivery outcomes.

3. Methodology

The project followed a **systematic and structured analytical workflow**, designed to transform the raw logistics dataset into meaningful business insights. The entire analysis was carried out using **Microsoft Excel**, which served as the primary tool for data pre-processing, analysis and visualization.

The methodological framework consisted of six major stages - starting from data import and cleaning, followed by feature creation, statistical analysis, visualization through pivot tables and charts and ending with an interactive dashboard for result interpretation.

Each stage was carefully executed to ensure accuracy, consistency and analytical depth.

3.1 Workflow Overview

The complete process is summarized below and visually represented in **Figure 3.1.1:**

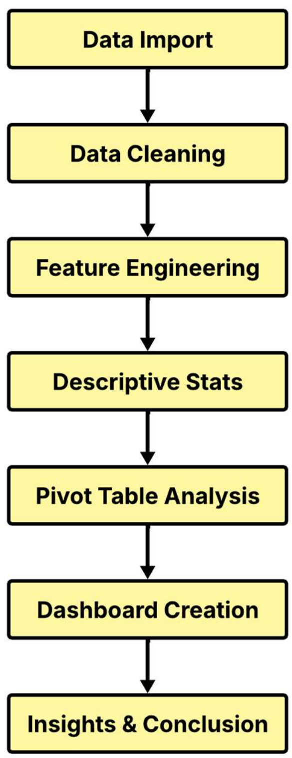


Figure 3.1.1: *Project Workflow Diagram*

3.2 Step-by-Step Summary

Step 1: Data Import and Familiarization

- The dataset was imported from **Kaggle** and opened in Excel. An initial review was performed to understand its structure, types of variables and potential data quality issues.
- Columns such as **Order_Date**, **Agent_Age** and **Delivery_Time** were identified as key fields for analysis.

Step 2: Data Cleaning and Pre-processing

- This stage involved **removing duplicates**, **handling missing** or **null values**, **formatting date/time fields** correctly and **validating** numerical entries.
- **Conditional formatting** was used to highlight anomalies and logical formulas were applied to correct inconsistencies (e.g. negative waiting times).
- This ensured that the dataset was accurate and reliable before analysis.

Step 3: Feature Engineering

- New variables were created from existing columns to add analytical depth.
- These included **Day_Name**, **Wait_Time_Min**, **Order_Hour**, **Hour_Bucket**, **Agent_Age_Bucket** and **Distance_Km**.
- Each of these derived columns played a role in answering specific business questions such as analyzing delivery performance across time, distance and demographics.

Step 4: Descriptive Statistics and Analysis

- Descriptive statistics were calculated using Excel's **Data Analysis ToolPak** to summarize numerical fields.
- Measures like **mean**, **median**, **standard deviation** and **range** helped in understanding the **central tendency** and spread of **delivery times**, **agent ratings** and **distances**.
- This stage provided the foundation for detecting trends and variations.

Step 5: Pivot Table Analysis and Visualization

- **Pivot tables** were created to summarize data across multiple dimensions such as **weather**, **traffic**, **area** and **vehicle type**.
- Each pivot helped answer one or more business questions. From these tables, pivot charts including bar, column, pie and scatter plots were generated to visualize performance and identify relationships between variables.

Step 6: Dashboard Creation

- The final stage was the development of an **interactive Excel dashboard**.
- **Key Performance Indicators** (KPIs) such as **Average Delivery Time**, **Average Wait Time**, **Average Rating** and **Total Orders** were displayed at the top.
- Charts were arranged below them in a clean layout and slicers were added for dynamic filtering by **Weather**, **Traffic**, **Area** and **Vehicle**.
- This made the results easily interpretable and visually appealing.

Step 7: Insights and Interpretation

- The trends identified from statistical summaries and visualizations were consolidated into key insights.
- Relationships such as **distance vs. delivery time**, **weather vs. waiting time** and **agent rating vs. age** were analyzed to understand operational dependencies.
- These insights were later summarized in the conclusion to highlight opportunities for improving delivery efficiency.

4. Data Cleaning and Pre-processing

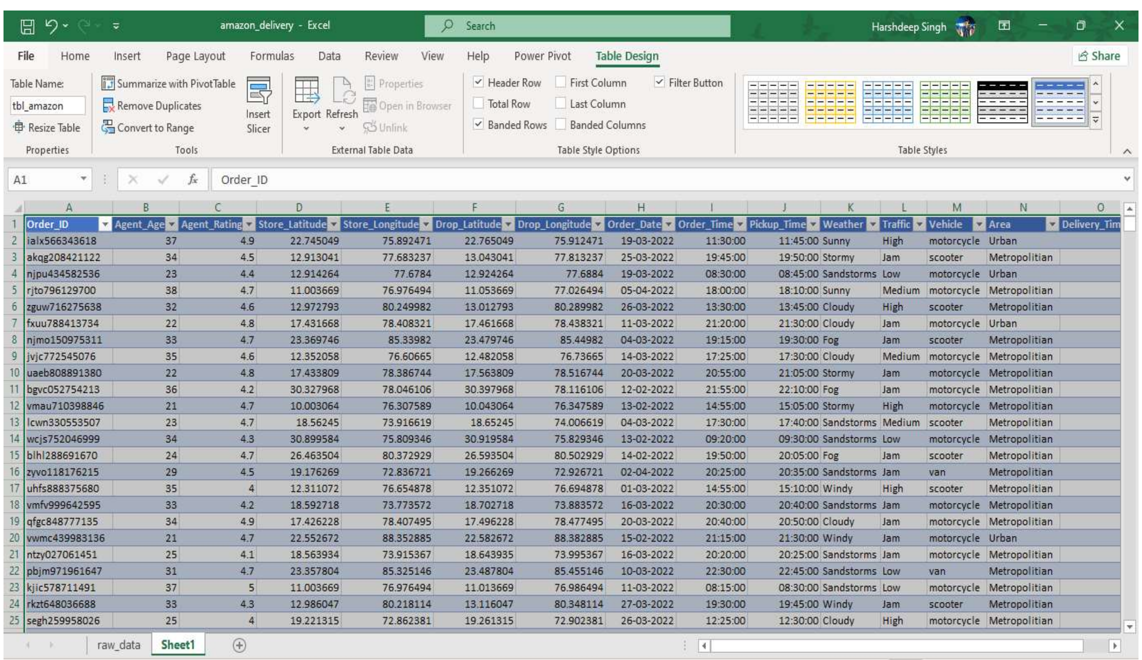
Data cleaning and pre-processing formed the foundation of this project. Since the dataset obtained from Kaggle contained real-world logistics information, it included several irregularities such as inconsistent formats, missing values and outliers. To ensure accurate and meaningful analysis, the raw data was cleaned, validated and standardized using **Microsoft Excel**.

This stage was crucial, as errors or inconsistencies in data can lead to misleading conclusions. Therefore, each column was examined carefully and systematic cleaning operations were performed to make the dataset reliable and analysis-ready.

4.1 Converting Raw Data into a Structured Table

The first step in pre-processing was to convert the raw data into an Excel **structured table** using the shortcut `Ctrl + T`. This enabled automatic filtering, sorting and structured references in formulas (e.g., using `[@[Order_Time]]` instead of cell references like C2).

The table was named **tbl_amazon** to simplify referencing in formulas used throughout the analysis.



Order_ID	Agent_Age	Agent_Rating	Store_Latitude	Store_Longitude	Drop_Latitude	Drop_Longitude	Order_Date	Order_Time	Pickup_Time	Weather	Traffic	Vehicle	Area	Delivery_Tim
ialx566343618	37	4.9	22.745049	75.892471	22.765049	75.912471	19-03-2022	11:30:00	11:45:00	Sunny	High	motorcycle	Urban	
akqg208421122	34	4.5	12.913041	77.683237	13.043041	77.813237	25-03-2022	19:45:00	19:50:00	Stormy	Jam	scooter	Metropolitian	
njpu434582536	23	4.4	12.914264	77.6784	12.924264	77.6884	19-03-2022	08:30:00	08:45:00	Sandstorms	Low	motorcycle	Urban	
rjto796129700	38	4.7	11.003669	76.976494	11.053669	77.026494	05-04-2022	18:00:00	18:10:00	Sunny	Medium	motorcycle	Metropolitian	
zguw716275638	32	4.6	12.972793	80.249982	13.012793	80.289982	26-03-2022	13:30:00	13:45:00	Cloudy	High	scooter	Metropolitian	
fxuu788413734	22	4.8	17.431668	78.408321	17.461668	78.438321	11-03-2022	21:20:00	21:30:00	Cloudy	Jam	motorcycle	Urban	
njmo150975311	33	4.7	23.369746	85.33982	23.479746	85.44982	04-03-2022	19:15:00	19:30:00	Fog	Jam	scooter	Metropolitian	
jvjc772545076	35	4.6	12.352058	76.60665	12.482058	76.73665	14-03-2022	17:25:00	17:30:00	Cloudy	Medium	motorcycle	Metropolitian	
uaeb808891380	22	4.8	17.433809	78.386744	17.563809	78.516744	20-03-2022	20:55:00	21:05:00	Stormy	Jam	motorcycle	Metropolitian	
bgvc052754213	36	4.2	30.327968	78.046106	30.397968	78.116106	12-02-2022	21:55:00	22:10:00	Fog	Jam	motorcycle	Metropolitian	
vmau710398846	21	4.7	10.003064	76.307589	10.043064	76.347589	13-02-2022	14:55:00	15:05:00	Stormy	High	motorcycle	Metropolitian	
lcwn330553507	23	4.7	18.56245	73.916619	18.65245	74.006619	04-03-2022	17:30:00	17:40:00	Sandstorms	Medium	scooter	Metropolitian	
wcjs752046999	34	4.3	30.899584	75.809346	30.919584	75.829346	13-02-2022	09:20:00	09:30:00	Sandstorms	Low	motorcycle	Metropolitian	
blhl288691670	24	4.7	26.463504	80.372929	26.593504	80.502929	14-02-2022	19:50:00	20:05:00	Fog	Jam	scooter	Metropolitian	
zyvo118176215	29	4.5	19.176269	72.836721	19.266269	72.926721	02-04-2022	20:25:00	20:35:00	Sandstorms	Jam	van	Metropolitian	
uhfs888375680	35	4	12.311072	76.654878	12.351072	76.694878	01-03-2022	14:55:00	15:10:00	Windy	High	scooter	Metropolitian	
vmfv999642595	33	4.2	18.592718	73.773572	18.702718	73.883572	16-03-2022	20:30:00	20:40:00	Sandstorms	Jam	motorcycle	Metropolitian	
qfgc848777135	34	4.9	17.426228	78.407495	17.496228	78.477495	20-03-2022	20:40:00	20:50:00	Cloudy	Jam	motorcycle	Metropolitian	
vwmc439983136	21	4.7	22.552672	88.352885	22.582672	88.382885	15-02-2022	21:15:00	21:30:00	Windy	Jam	motorcycle	Urban	
ntzy027061451	25	4.1	18.563934	73.915367	18.643935	73.995367	16-03-2022	20:20:00	20:25:00	Sandstorms	Jam	motorcycle	Metropolitian	
pbjm971961647	31	4.7	23.357804	85.325146	23.487804	85.455146	10-03-2022	22:30:00	22:45:00	Sandstorms	Low	van	Metropolitian	
kjic578711491	37	5	11.003669	76.976494	11.013669	76.986494	11-03-2022	08:15:00	08:30:00	Sandstorms	Low	motorcycle	Metropolitian	
rkzt648036688	33	4.3	12.986047	80.218114	13.116047	80.348114	27-03-2022	19:30:00	19:45:00	Windy	Jam	scooter	Metropolitian	
segh259958026	25	4	19.221315	72.862381	19.261315	72.902381	26-03-2022	12:25:00	12:30:00	Cloudy	High	motorcycle	Metropolitian	

***Figure 4.1.1**: Dataset converted into a structured table using "Ctrl + T"*

4.2 Formatting and Validating Data Types

Each column in the dataset was checked to ensure that it had the correct data type:

- Columns such as **Order_Date** and **Order_Time**, **Pickup_Time** were formatted as **Date** and **Time** respectively.
- Columns like **Agent_Age**, **Agent_Rating** and **Delivery_Time** were formatted as **Number**.
- Categorical fields such as **Weather**, **Traffic**, **Area** and **Vehicle** were formatted as **Text**.

This ensured that all subsequent calculations and functions in Excel returned accurate results.

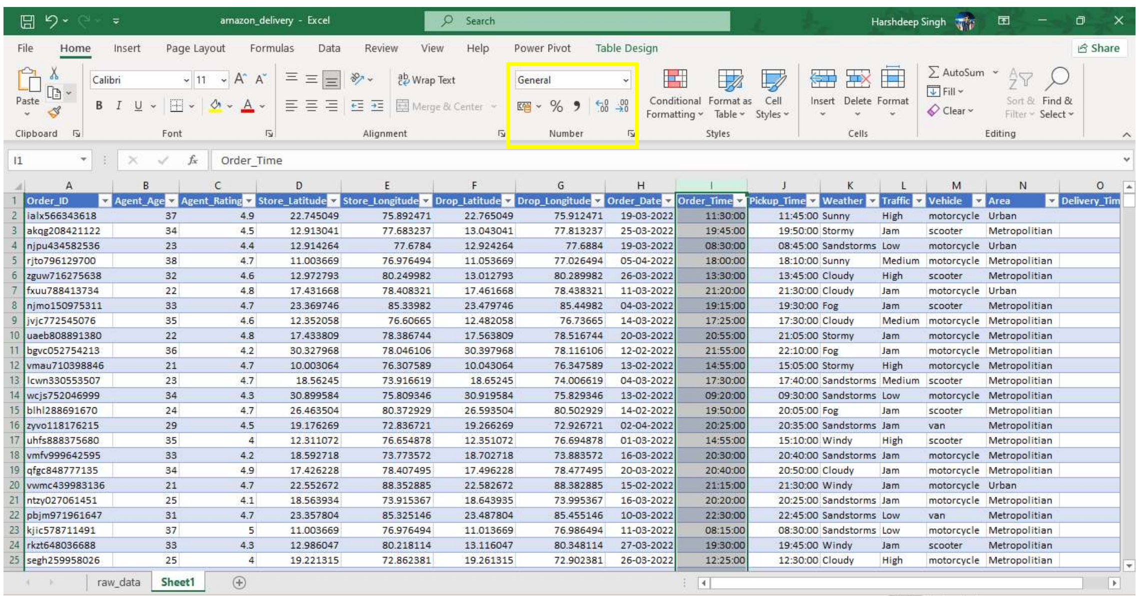


Figure 4.2.1: *The datatype of the column Order_Time was General so changed it into Time*

4.3 Detecting and Handling Missing Values

Missing or null entries were identified using the `COUNTBLANK()` function and **conditional formatting** (highlighting empty cells in red color). The following strategies were applied to handle incomplete data:

- **Order_Time / Pickup_Time:** Rows with missing or invalid time entries were removed, as they prevented accurate waiting-time computation.
- **Zero Values:** Any zero entries in key fields like **Agent_Age** or **Distance** were reviewed and corrected if inconsistent.

- **Agent_Rating:** Missing values were replaced with the mean rating using `=IF(ISBLANK([@[Agent_Rating]]),AVERAGE([Agent_Rating]),[@[Agent_Rating]])`.

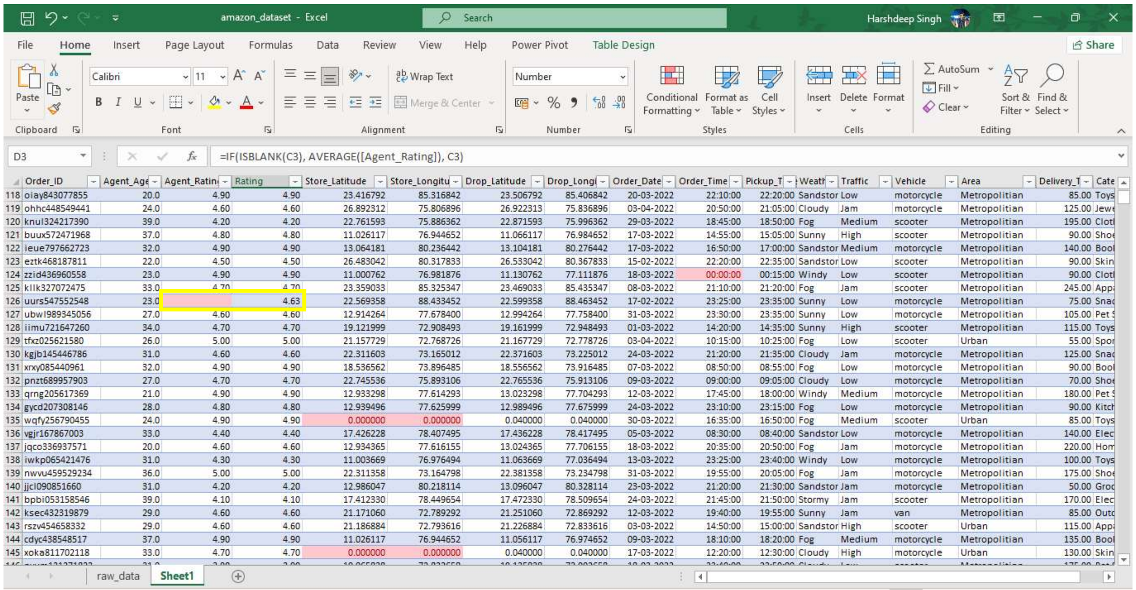


Figure 4.3.1: *Filling up the detected missing values using formulas*

4.4 Removing Duplicates and Ensuring Uniqueness

To verify data integrity, the **Order_ID** column was used to check for duplicate records using Excel's **Remove Duplicates** feature.

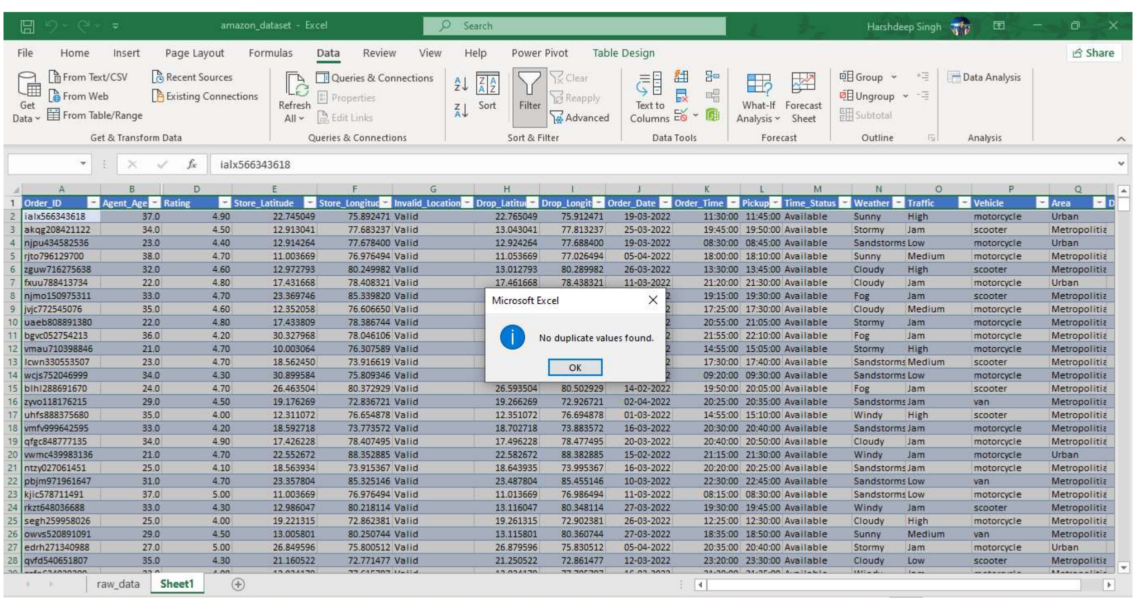


Figure 4.4.1: *No duplicate Order_ID entries detected*

4.5 Handling NaN Values

Detected **NaN values** in the **Order_Time** column which could lead to inconsistencies in the further analysis. There were **91 entries** out of the **43739 rows** which is nearly about **2.5%** so I decided to delete these rows.

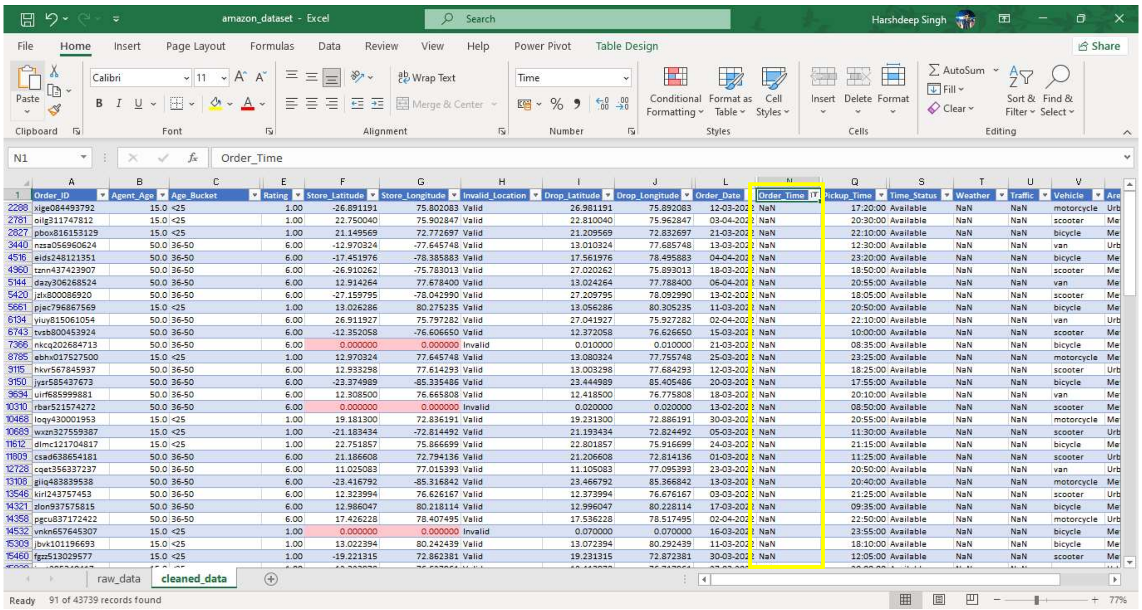


***Figure 4.5.1**: NaN Values in the Order_Time Column*

4.6 Final Data Validation and Export

Once all corrections were made, the dataset was thoroughly reviewed to ensure:

- No blank cells remained in critical columns.
- Date and time calculations produced logical results.
- Duplicate records were completely removed.
- No invalid numeric values persisted.

After validation, the cleaned dataset was saved as **amazon_dataset.xlsx** and became the foundation for further analysis and visualization.

5. Feature Engineering

After the dataset was cleaned and validated, the next step was to enhance it through **feature engineering**. Feature engineering refers to the process of creating new, meaningful variables from existing data to improve analytical depth and discover hidden patterns.

In this project, several new columns were derived using Excel formulas to help answer the business questions defined earlier such as analyzing delivery patterns across time, geography and agent demographics.This process not only improved the interpretability of the dataset but also allowed for more insightful visualizations and performance comparisons across different conditions.

5.1 Rationale Behind Feature Engineering

The original dataset, although rich in attributes, lacked direct indicators of operational efficiency such as **waiting time**, **time of day**, **delivery distance**. Hence, additional columns were created to:

- Identify temporal trends in orders and pickups.
- Categorize agents based on age and performance.
- Estimate distances using geographical coordinates.
- Group deliveries into time buckets for pattern observation.

5.2 Derived Columns and Their Purpose

New Column	Formula	Purpose
Age_Bucket	`=IF([@[Agent_Age]]<25,"<25",IF([@[Agent_Age]]<=35,"25-35",IF([@[Agent_Age]]<=50,"36-50","50+")))`	Groups delivery agents into **age ranges** for performance and demographic insights.
Distance_Km	`=12742*ASIN(SQRT(SIN(RADIANS([@[Drop_Latitude]] - [@[Store_Latitude]])/2)^2 + COS(RADIANS([@[Store_Latitude]])) * COS(RADIANS([@[Drop_Latitude]])) * SIN(RADIANS([@[Drop_Longitude]] - [@[Store_Longitude]])/2)^2))`	Computes the **straight-line (great-circle) distance** in kilometers between store and drop locations using the **Haversine formula**.

Day_Name	`=TEXT([@[Order_Date]], "dddd")`	Extracts the **day of the week** from each order date to analyze daily order patterns.
Order_Hour	`=HOUR([@[Order_Time]])`	Extracts the **hour** from the order timestamp to study hourly order trends.
Hour_Bucket	`=IF([@[Order_Hour]]<8, "Early Morning", IF([@[Order_Hour]]<12, "Morning", IF([@[Order_Hour]]<16, "Afternoon", IF([@[Order_Hour]]<20, "Evening", "Night"))))`	Categorizes orders into **time-of-day segments** (Morning, Afternoon, etc.) for behavioral or operational analysis.
Wait_Time_Min	`=IF([@[Pickup_Time]]<[@[Order_Time]],(([@[Pickup_Time]]+1)-[@[Order_Time]])*1440,([@[Pickup_Time]]-[@[Order_Time]])*1440)`	Calculates the **waiting time (in minutes)** between order placement and pickup. 1440 converts Excel's time fraction (days) into minutes.

5.3 Example Formula Demonstration

To represent the concept of feature engineering in the report visually, one of the derived columns - **Wait_Time_Min** is highlighted as an example.

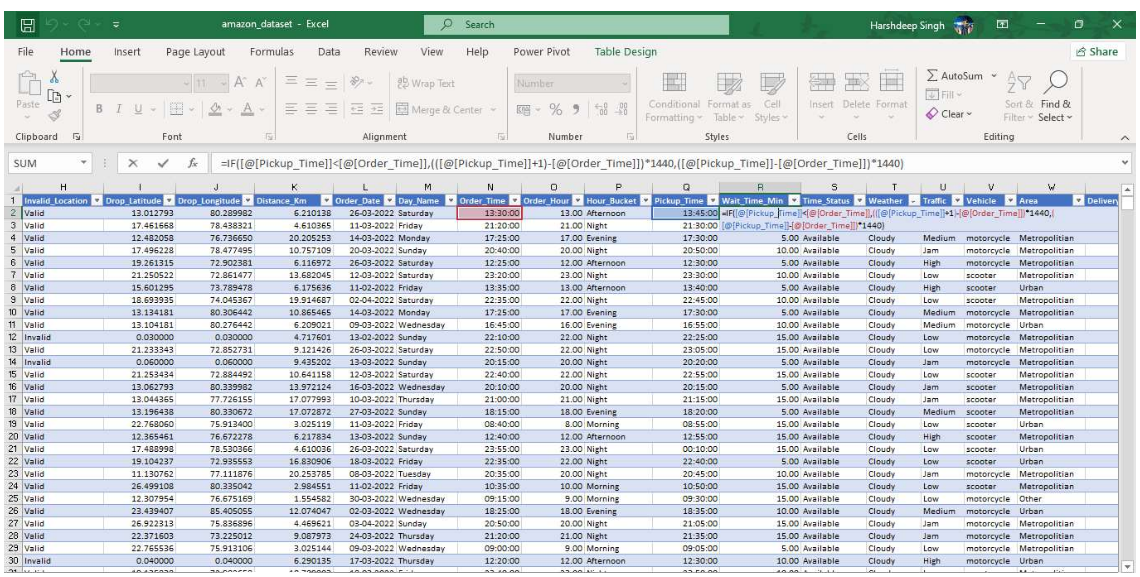


Figure 5.3.1: *Formula demonstration for derived column "Wait_Time_Min" showing relationship between Order_Time and Pickup_Time.*

5.4 Validation of Engineered Features

After the new columns were created, multiple validation checks were conducted:

- **Data Type Verification:** Ensured that derived columns were correctly formatted (e.g., numeric for distances, text for buckets).
- **Error Checking:** The `IFERROR()` function was temporarily applied to identify invalid computations or missing data dependencies.
- **Logical Testing:** Relationships were validated for example **-** longer distances generally resulting in higher delivery times and peak hours corresponding to higher order volumes.

This verification ensured that every new feature contributed to accurate and meaningful insights.

5.5 Outcome of Feature Engineering

After completing this step, the dataset grew both in **analytical depth** and **interpretability**. The new columns enabled:

- Detailed **time-based**, **distance-based** and **agent-based** analyses.
- Richer pivot table creation for comparative studies.
- More interactive and layered insights in the final dashboard.

Feature engineering thus transformed a raw logistics dataset into a structured, business-ready dataset capable of answering the predefined analytical questions effectively.

6. Descriptive Statistics and Analysis

After the dataset was cleaned and enhanced through feature engineering, the next phase of the analysis involved performing **Descriptive Statistics** to summarize and understand the distribution, central tendencies and spread of key numerical variables. Descriptive statistics form the foundation of data analytics by converting large volumes of raw data into meaningful summary measures that highlight general trends and variations.

The analysis in this phase was performed using **Microsoft Excel's Data Analysis ToolPak**, as well as Excel formulas such as `AVERAGE()`, `MEDIAN()`, `STDEV()` and `MAX()` to cross-verify results manually.

6.1 Purpose of Descriptive Statistics

The purpose of this step was to:

- Understand the **average behavior** of key metrics such as delivery time, waiting time, agent rating and delivery distance.
- Identify **variability** within the dataset using measures like range and standard deviation.
- Detect potential **outliers or extreme observations**.
- Prepare a statistical foundation for deeper analysis through pivot tables and visualizations.

6.2 Variables Analyzed

The following key numerical columns were analyzed using descriptive statistics:

- Delivery_Time
- Wait_Time_Min
- Agent_Rating
- Agent_Age
- Distance_Km

These variables were selected because they directly represent key performance indicators (KPIs) in delivery logistics.

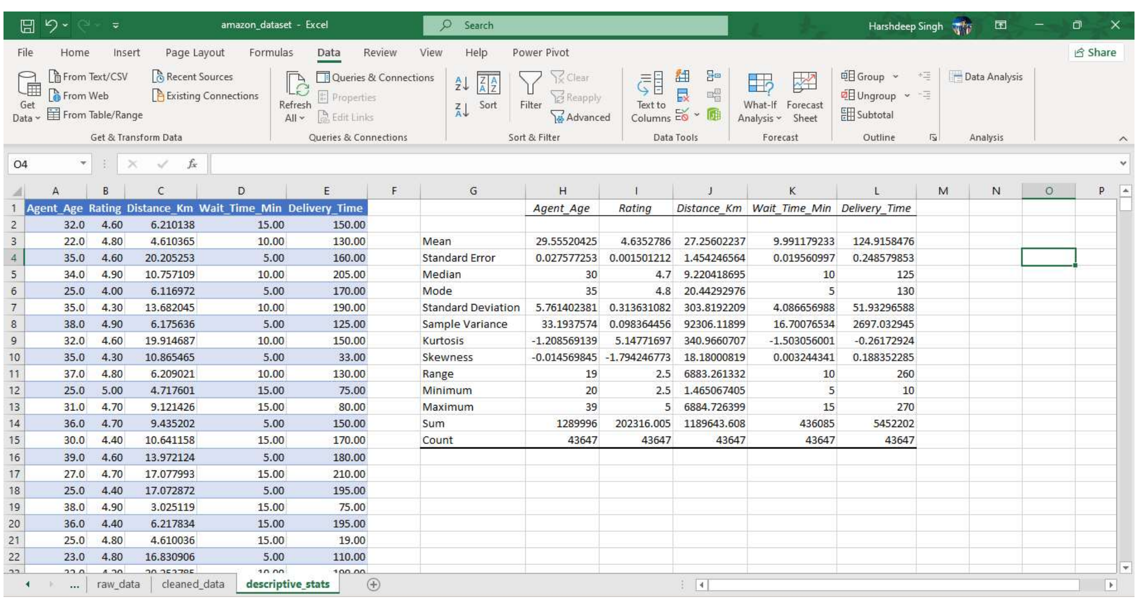
6.3 Statistical Measures Computed

The following descriptive measures were computed for each numeric variable using the **Data Analysis ToolPak**:

Statistic	**Interpretation**
Mean (Average)	Indicates the central tendency of data, representing the typical value.
Median	Represents the middle value, unaffected by extreme values.
Mode	Identifies the most frequently occurring value.
Standard Deviation (Stdev)	Measures the dispersion or variability of the data.
Range (Max-Min)	Highlights the spread between the highest and lowest values.
Count	Shows the total number of valid observations for the variable.
Sum	The total aggregate of a numerical field

6.4 Output and Observations

The **Data Analysis ToolPak** was used (accessible via *Data → Data Analysis → Descriptive Statistics*) to generate summary statistics for all key numeric variables. The output generated was placed on a new worksheet titled **"descriptive_stats"** for easy reference.



Agent_Age	Rating	Distance_Km	Wait_Time_Min	Delivery_Time
32.0	4.60	6.210138	15.00	150.00
22.0	4.80	4.610365	10.00	130.00
35.0	4.60	20.205253	5.00	160.00
34.0	4.90	10.757109	10.00	205.00
25.0	4.00	6.116972	5.00	170.00
35.0	4.30	13.682045	10.00	190.00
38.0	4.90	6.175636	5.00	125.00
32.0	4.60	19.914687	10.00	150.00
35.0	4.30	10.865465	5.00	33.00
37.0	4.80	6.209021	10.00	130.00
25.0	5.00	4.717601	15.00	75.00
31.0	4.70	9.121426	15.00	80.00
36.0	4.70	9.435202	5.00	150.00
30.0	4.40	10.641158	15.00	170.00
39.0	4.60	13.972124	5.00	180.00
27.0	4.70	17.077993	15.00	210.00
25.0	4.40	17.072872	5.00	195.00
38.0	4.90	3.025119	15.00	75.00
36.0	4.40	6.217834	15.00	195.00
25.0	4.80	4.610036	15.00	19.00
23.0	4.80	16.830906	5.00	110.00

	Agent_Age	Rating	Distance_Km	Wait_Time_Min	Delivery_Time
Mean	29.55520425	4.6352786	27.25602237	9.991179233	124.9158476
Standard Error	0.027577253	0.001501212	1.454246564	0.019560997	0.248579853
Median	30	4.7	9.220418695	10	125
Mode	35	4.8	20.44292976	5	130
Standard Deviation	5.761402381	0.313631082	303.8192209	4.086656988	51.93296588
Sample Variance	33.1937574	0.098364456	92306.11899	16.70076534	2697.032945
Kurtosis	-1.208569139	5.14771697	340.9660707	-1.503056001	-0.26172924
Skewness	-0.014569845	-1.794246773	18.18000819	0.003244341	0.188352285
Range	19	2.5	6883.261332	10	260
Minimum	20	2.5	1.465067405	5	10
Maximum	39	5	6884.726399	15	270
Sum	1289996	202316.005	1189643.608	436085	5452202
Count	43647	43647	43647	43647	43647

***Figure 6.1:** Output of Descriptive Statistics generated using Excel Data Analysis ToolPak.*

From the statistical output, the following key observations were made:

1. Delivery_Time

- The mean delivery time was found to be moderate, with a noticeable spread, indicating operational variability across orders.
- A few orders recorded significantly higher delivery times, likely caused by heavy traffic or long delivery distances.

2. Wait_Time_Min

- The average waiting time between order placement and pickup was relatively low, but the presence of high standard deviation suggested irregular pickup efficiency across time periods.
- Negative waiting times detected earlier were corrected, ensuring logical results.

3. Agent_Rating

- The mean agent rating was high (around 4.6), indicating generally good performance and customer satisfaction.
- Ratings showed little variation, suggesting that most agents maintained consistent service levels.

4. Agent_Age

- The age distribution showed a concentration of agents between 25 and 35 years, which was also reflected in the **Agent_Age_Bucket** feature.

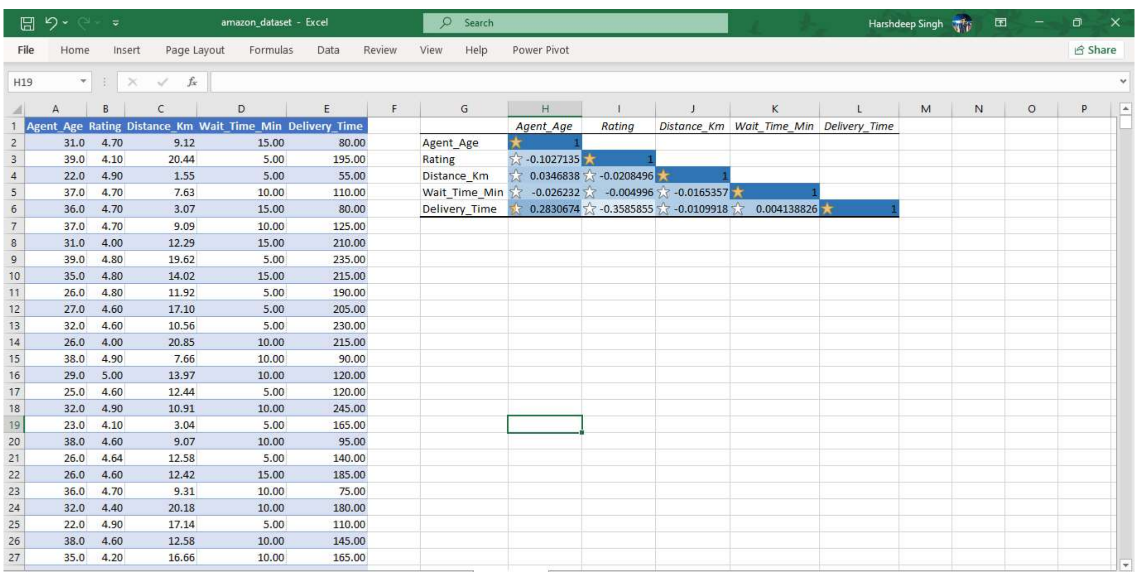
5. Distance_Km

- The average delivery distance was moderate, with some long-distance deliveries extending significantly beyond the mean.
- Standard deviation indicated variation between urban and semi-urban delivery regions.

6.5 Correlation Overview (Preliminary Analysis)

Before building visualizations, a **correlation matrix** was also generated between key numeric variables using **Data Analysis ToolPak** (accessible via *Data → Data Analysis → Correlation*). This helped identify relationships that could be explored later through pivot charts.

Variable Pair	**Expected Relationship**	**Observation**
Distance_Km vs. Delivery_Time	Positive	Longer distances generally resulted in higher delivery times.
Traffic vs. Delivery_Time	Positive (Categorical-Numeric)	High traffic density corresponded to increased delivery durations.
Wait_Time_Min vs. Delivery_Time	Slightly Positive	Longer waiting times occasionally led to longer total deliveries.
Agent_Rating vs. Delivery_Time	Slightly Negative	Higher-rated agents tended to deliver faster and more efficiently.



Agent_Age	Rating	Distance_Km	Wait_Time_Min	Delivery_Time
31.0	4.70	9.12	15.00	80.00
39.0	4.10	20.44	5.00	195.00
22.0	4.90	1.55	5.00	55.00
37.0	4.70	7.63	10.00	110.00
36.0	4.70	3.07	15.00	80.00
37.0	4.70	9.09	10.00	125.00
31.0	4.00	12.29	15.00	210.00
39.0	4.80	19.62	5.00	235.00
35.0	4.80	14.02	15.00	215.00
26.0	4.80	11.92	5.00	190.00
27.0	4.60	17.10	5.00	205.00
32.0	4.60	10.56	5.00	230.00
26.0	4.00	20.85	10.00	215.00
38.0	4.90	7.66	10.00	90.00
29.0	5.00	13.97	10.00	120.00
25.0	4.60	12.44	5.00	120.00
32.0	4.90	10.91	10.00	245.00
23.0	4.10	3.04	5.00	165.00
38.0	4.60	9.07	10.00	95.00
26.0	4.64	12.58	5.00	140.00
26.0	4.60	12.42	15.00	185.00
36.0	4.70	9.31	10.00	75.00
32.0	4.40	20.18	10.00	180.00
22.0	4.90	17.14	5.00	110.00
38.0	4.60	12.58	10.00	145.00
35.0	4.20	16.66	10.00	165.00

	Agent_Age	Rating	Distance_Km	Wait_Time_Min	Delivery_Time
Agent_Age	1				
Rating	-0.1027135	1			
Distance_Km	0.0346838	-0.0208496	1		
Wait_Time_Min	-0.026232	-0.004996	-0.0165357	1	
Delivery_Time	0.2830674	-0.3585855	-0.0109918	0.004138826	1

Figure 6.5.1: *Correlation matrix and Conditional Formatting to highlight relation*

6.6 Interpretation

The descriptive and correlation analysis provided the initial statistical foundation for the rest of the project. It confirmed that:

- Operational factors such as **traffic**, **weather** and **distance** play significant roles in influencing delivery time.

- Human factors like **agent experience** and **performance** also contribute to operational consistency.
- There exists measurable variability across different regions and time slots, justifying further visual exploration using pivot tables and dashboards.
- A weak positive correlation (r = 0.2) was observed between **Agent_Age** and **Delivery_Time**, indicating that slightly older delivery agents tended to have longer delivery durations on average. However, the relationship was not strong enough to conclude that age alone determines performance.

7. Pivot Table Analysis and Visualization

After performing descriptive statistical analysis, the next step was to **visually explore the dataset** using **Pivot Tables** and **Charts** in Microsoft Excel. Pivot tables served as a powerful tool for summarizing large volumes of data across multiple dimensions, allowing for pattern recognition and insight generation in a simple yet effective manner.

This stage helped transform numerical summaries into actionable visual representations, directly addressing the business questions framed at the start of the project.

7.1 Purpose of Pivot Table Analysis

The primary purpose of using pivot tables was to:

- Identify relationships between delivery performance and various factors such as traffic, weather and distance.
- Analyze trends across **time**, **area** and **agent demographics**.
- Create dynamic and interactive visualizations that could later be used in the **dashboard**.
- Summarize metrics such as average delivery time, waiting time and agent ratings for better interpretability.

Pivot tables also allowed filtering and grouping of data in multiple ways, which made them ideal for identifying both micro and macro patterns in delivery behavior.

7.2 Creation of Pivot Tables

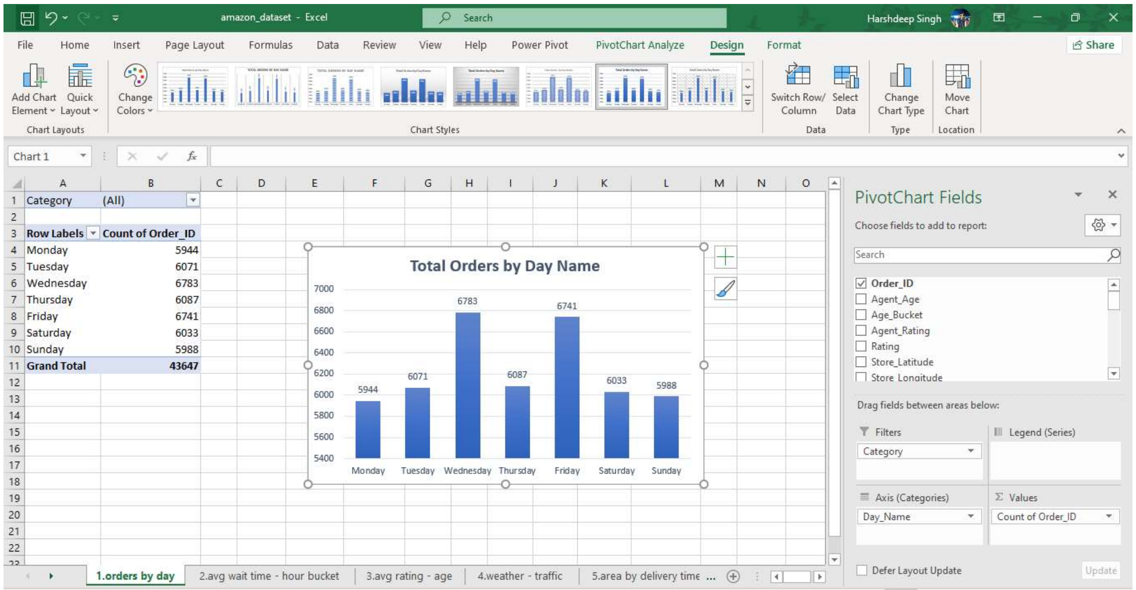
Multiple pivot tables were created each focusing on a specific aspect of logistics performance. All pivots were based on the cleaned and feature-engineered dataset.

The process followed for creating a pivot table was as below:

1. Selecting the complete table (**tbl_Amazon**) → **Insert** → **Pivot Table**.
2. Choosing “**New Worksheet**” as the output location for clarity.
3. Dragging and dropping relevant fields into **Rows**, **Columns**, **Values** and **Filters** areas.
4. Summarizing numerical fields using `AVERAGE`, `COUNT`, or `SUM` functions depending on the business question.
5. Inserted best corresponding Chart for Data Visualization.

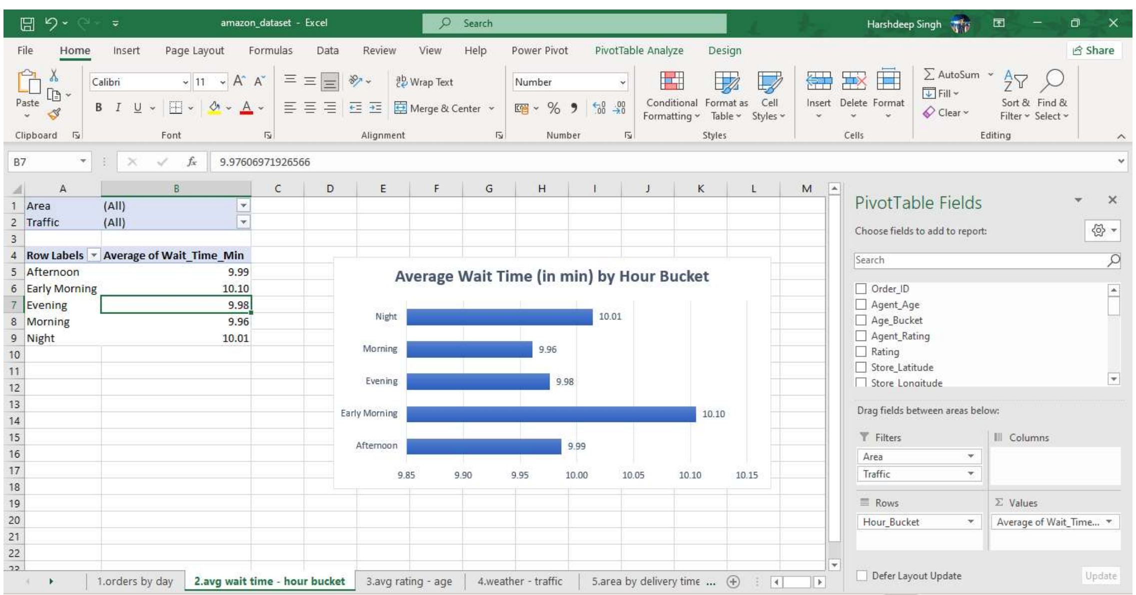
7.3 Key Pivot Tables and Charts Created

1. Orders by Day



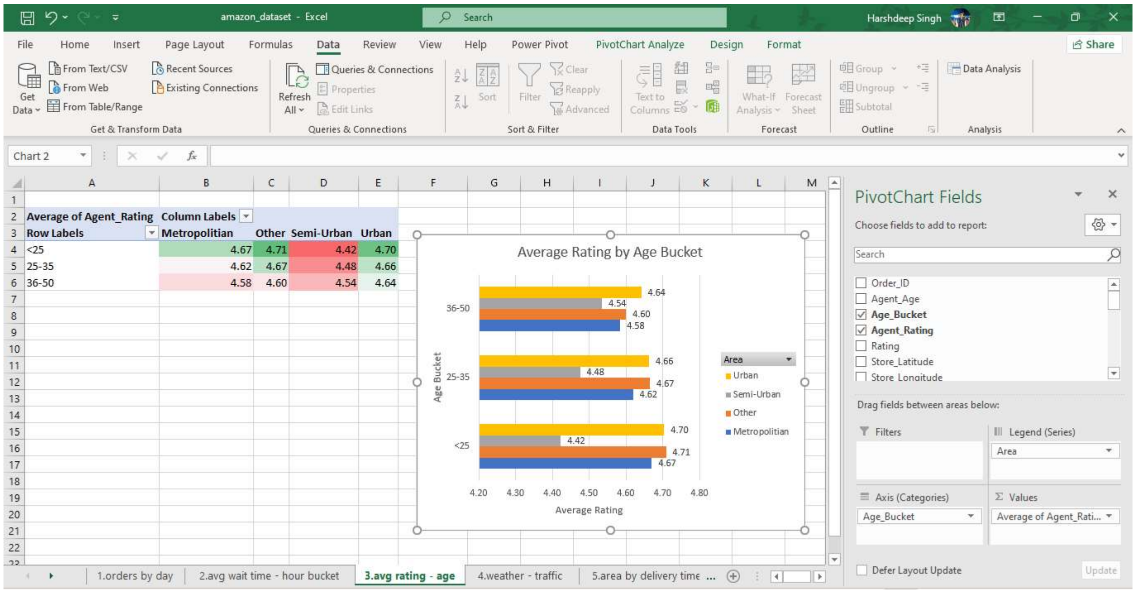
The analysis reveals that **Wednesday recorded the highest number of orders**, followed closely by **Friday** and **Thursday**.

2. Average Wait Time by Hour Bucket



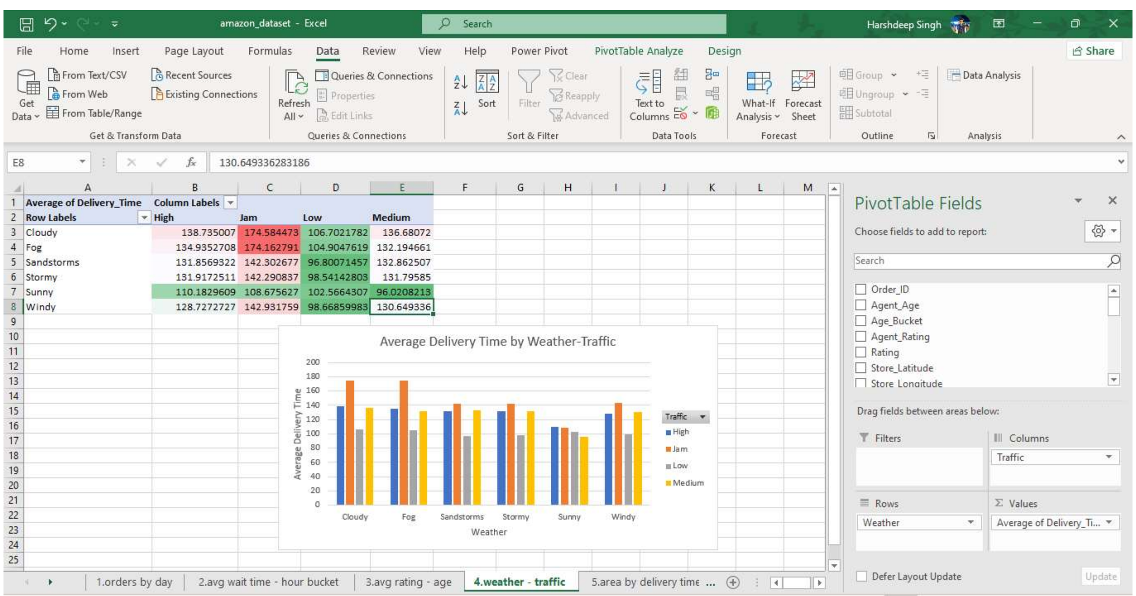
The analysis shows that the **average waiting time is highest during the Early Morning and Night hours**, while it remains comparatively lower during the daytime. Used area and traffic in filters to see further breakdown of waiting time.

3. Average Rating by Agent Age Bucket



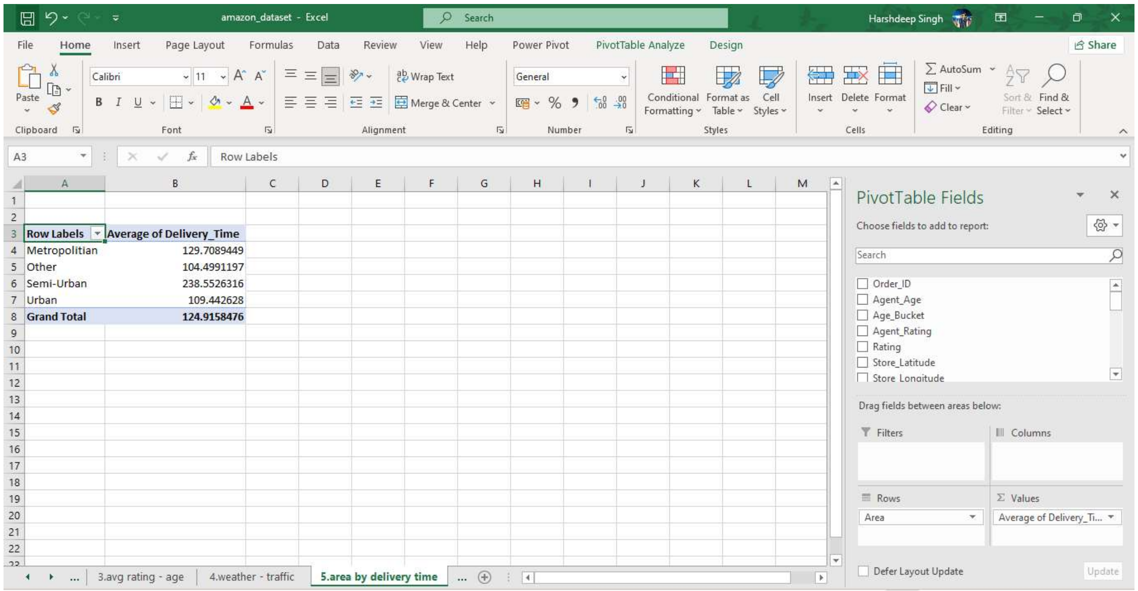
Delivery agents aged **below 25 years** maintained high ratings across most areas, except in **semi-urban regions**, where the **36-50 age group** performed best. **Conditional formatting** was used to highlight these differences, with **green for highest** and **red for lowest** ratings.

4. Delivery time by weather and traffic



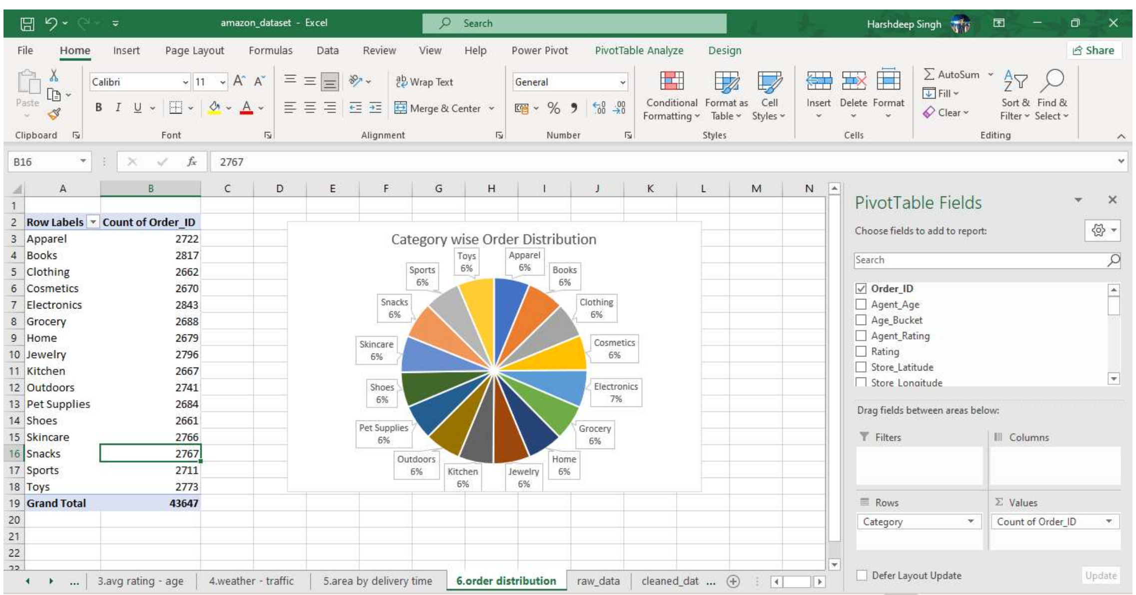
The **average delivery time increases significantly** during **foggy or cloudy weather** combined with **jammed traffic** conditions. **Conditional formatting** highlights this pattern clearly - with **red cells** indicating the highest delivery delays.

5. Average Delivery Time by Area



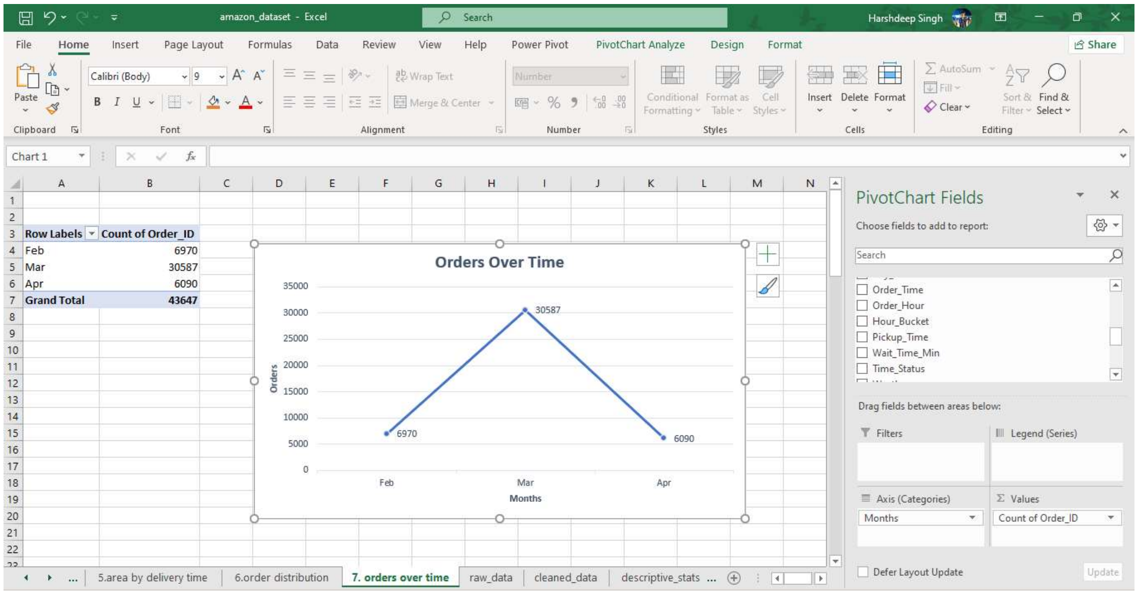
The **average delivery time is highest in Semi-Urban areas**, followed by **Metropolitan regions**. This suggests that route complexity or infrastructure limitations in semi-urban zones may contribute to longer delivery durations.

6. Category Wise Distribution



The **Electronics category** recorded the **highest number of orders**, contributing roughly **7% of total deliveries**. This indicates strong customer demand and higher order frequency for electronic products compared to other categories.

7. Orders over time



The analysis shows that **March recorded the highest number of orders**, indicating a clear peak in delivery activity during that month.

7.4 Linking Pivots with Slicers (Preparation for Dashboard)

Before building the final dashboard, slicers were tested and linked to each pivot table. This allowed interactive filtering of data by **Weather**, **Traffic**, **Vehicle** and **Area** ensuring that all charts updated dynamically based on slicer selections.

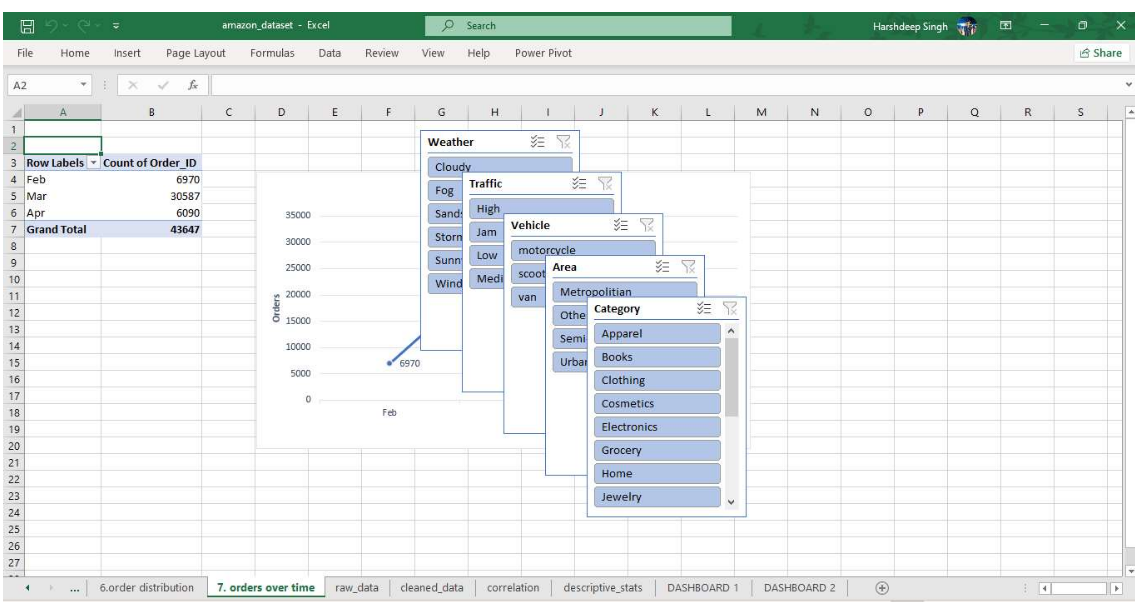


Figure 7.4.1: *Slicers Addition*

8. Dashboard Creation

After completing pivot table analysis and generating multiple visualizations, the next stage of the project involved consolidating all findings into a **single interactive dashboard**. The goal of the dashboard was to provide a clear, dynamic and visually appealing summary of key performance metrics related to Amazon's delivery operations.

The dashboard was designed using **Microsoft Excel's visualization tools**, such as pivot charts, slicers and KPI cards, to enable quick exploration of trends and relationships between operational factors like traffic, weather, agent ratings and delivery efficiency.

8.1 Purpose

The primary purpose of the dashboard was to:

- Present analytical results in an **interactive and easily interpretable format**.
- Provide a **management-style overview** of logistics performance using measurable KPIs.
- Enable **dynamic filtering** through slicers for deeper, condition-specific exploration.
- Combine multiple visual insights into one comprehensive report layout.

By using the dashboard, any user even without technical knowledge could explore how delivery time, agent rating and environmental factors interact in real-world logistics performance.

8.2 Dashboard Design and Layout

The dashboard was built on a **dedicated worksheet** named "Dashboard." Its layout followed a structured and visually balanced design:

Section	**Description**
Header Section	Displayed the project title "Amazon Logistics Delivery Performance Dashboard" with a clean heading and background color.
KPI Section (Top Row)	Contained key metrics such as Average Delivery Time, Average Wait Time, Average Agent Rating, Average Distance and Total

	Orders Delivered. These were displayed using large font and cell formatting for easy readability.
Chart Section (Middle)	Featured four major pivot charts showcasing delivery performance by Traffic, Weather, Hour Bucket and Area. Each chart provided visual comparison and trend identification.
Slicer Panel (Right Side)	Included slicers for Weather, Traffic, Area and Vehicle Type. These allowed users to filter all visuals simultaneously and interactively.

8.3 Key Performance Indicators (KPIs)

Five major KPIs were calculated and displayed prominently at the top of the dashboard. These indicators summarize the overall performance of the delivery network.

1. Average Delivery Tine
2. Average Wait Time
3. Average Agent Rating
4. Average Distance (Km)
5. Total Orders Delivered

8.4 Dashboard Output

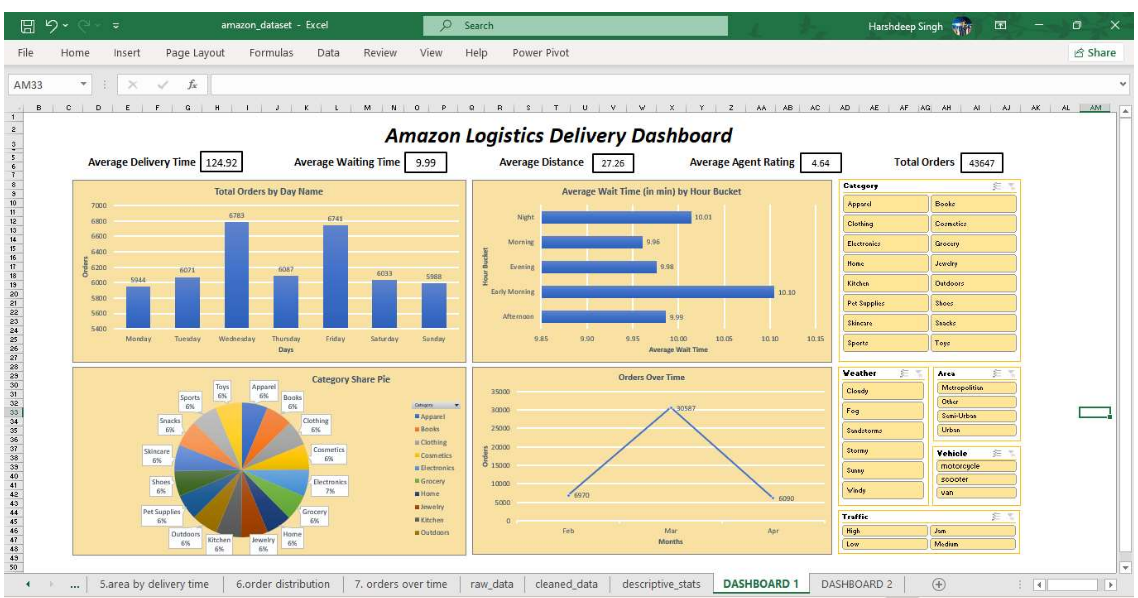


Figure 8.4.1: Dashboard Page-1

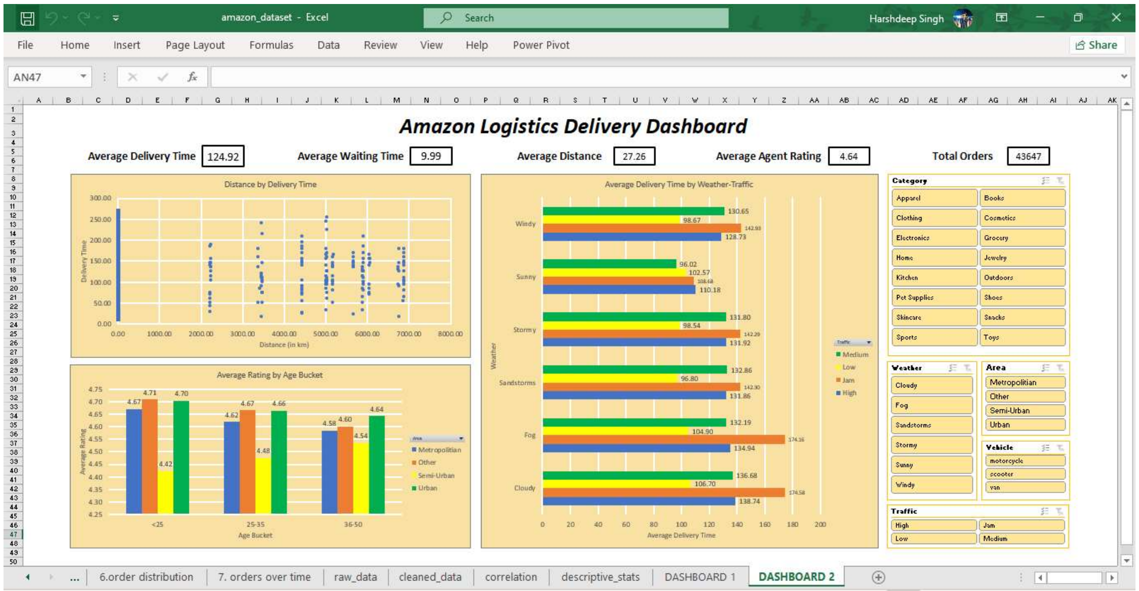


Figure 8.4.2: *Dashboard Page-2*

9. Insights

After completing data cleaning, feature engineering, statistical analysis and visualization, several key insights were derived from the Amazon Logistics dataset. The findings highlight how both **operational factors** (traffic, weather and distance) and **human factors** (agent age, rating and availability) influence delivery efficiency.

The discussion below summarizes the major insights obtained through descriptive statistics, pivot charts and the interactive dashboard.

9.1 Order Trends and Category Distribution

1. The analysis revealed that **Wednesday recorded the highest number of orders**, followed by **Friday** and **Thursday**, indicating a mid-week to end-week surge in customer activity.
2. Among product categories, **Electronics** showed the highest order count, contributing around **7% of total deliveries**.
3. This pattern suggests that demand planning and resource allocation should be focused around mid-week periods and high-demand categories.

9.2 Time-Based Waiting Patterns

1. Average waiting time was found to be **highest during early morning and night hours**, while **daytime deliveries** were generally faster.
2. When viewed by area and traffic, it was observed that **semi-urban and high-traffic regions** experienced longer waiting times, particularly during late hours.
3. This implies the need for better **shift scheduling** and **agent allocation** during off-peak periods.

9.3 Traffic and Weather Influence on Delivery Time

1. Deliveries conducted during **jammed traffic** and **foggy or cloudy weather** recorded the **longest delivery times**.
2. Traffic congestion and poor visibility were major external factors contributing to delays.
3. Such patterns highlight the importance of **traffic-aware routing** and **weather-based delivery forecasting** to maintain reliability in adverse conditions.

9.4 Area-Wise Delivery Performance

1. The **semi-urban regions** showed the **highest average delivery times**, followed by **metropolitan areas**, while **urban regions** performed comparatively better.
2. The variation is likely due to infrastructure differences and longer travel routes in semi-urban zones.
3. Improving **regional logistics planning** and **route optimization** could help balance these disparities.

9.5 Distance and Delivery Time Relationship

1. A clear **positive relationship** was observed between **distance** and **delivery time** longer distances naturally resulted in longer delivery durations.
2. However, a few short-duration long-distance deliveries indicated efficiency improvements due to better routing or agent performance.
3. This confirms that while distance is a primary factor, operational strategy also plays a crucial role in maintaining efficiency.

9.6 Agent Age and Performance

1. Agents aged **below 25 years** generally achieved **higher ratings**, showing strong overall performance and faster deliveries.
2. In **semi-urban areas**, agents aged **36-50 years** performed best, suggesting that experience helps navigate challenging routes more effectively.
3. This balance between youth and experience can be leveraged through **strategic agent assignments**.

9.7 Recommendations

Based on the findings, the following recommendations are proposed to improve delivery performance and operational planning in a realistic, data-driven manner:

1. **Adopt traffic-aware route optimization**

   Real-time traffic monitoring tools should be integrated into delivery planning. This can help agents automatically choose alternate routes during heavy congestion, especially in metropolitan areas where “Jammed” conditions caused up rise in delivery times.

2. **Plan proactively for adverse weather**

   Since cloudy and foggy conditions slowed deliveries, weather forecasts should be built into scheduling systems.

3. **Optimize shift patterns and agent scheduling**

   Waiting times were highest during early morning and night hours. Increasing the number of on-duty agents in these periods, or offering flexible shift rotations, can reduce idle time and ensure faster pickups for customers ordering outside regular hours.

4. **Strengthen semi-urban logistics coverage**

   Semi-urban regions showed the longest delivery durations, often due to inconsistent road networks and longer travel distances. Allocating additional resources, setting up small local hubs, or improving route familiarity training can help reduce average delivery times in these zones.

5. **Leverage agent strengths wisely**

   Younger agents (< 25 years) demonstrated quicker deliveries, while mid-aged agents (36-50 years) performed better in complex regions. Assigning routes based on each agent's strengths, speed, experience, local knowledge can create a balanced, efficient workforce.

6. **Anticipate seasonal and mid-week peaks**

   Order volumes peaked mid-week and during March, indicating predictable high-demand windows. Proactive staffing and vehicle readiness during these periods can help maintain service levels without overburdening teams.

7. **Maintain interactive performance dashboards**

   The dashboard built for this project can serve as a real-time monitoring tool. Regular updates with new data will help managers track delivery efficiency, agent performance and emerging trends, fostering a culture of continuous improvement.

10. Future Scope

The present analysis provided a detailed understanding of how operational, environmental and human factors influence delivery efficiency within Amazon's logistics network. However, the study can be expanded in several meaningful ways to gain even deeper insights and practical applications in the future.

1. **Integration of Machine Learning Techniques**

   The next step could involve **predictive modelling** to forecast delivery times, wait times, or traffic delays based on past patterns. Incorporating algorithms such as regression models or decision trees can transform this analysis from descriptive to predictive.

2. **Automation Using Power BI or Python**

   While this project was executed entirely in Excel, future work could employ **Power BI dashboards** or **Python-based analytics** to automate data refresh, enhance visualization and handle larger datasets more efficiently.

3. **Inclusion of Real-Time and Live Data Streams**

   Integrating **APIs** for live weather, traffic, or GPS tracking data could provide a more dynamic picture of ongoing deliveries, allowing for near real-time performance monitoring.

4. **Expanded Feature Engineering**

   Additional engineered variables such as **delivery speed** (km/min), **agent workload**, or **customer satisfaction** scores can provide new analytical dimensions and help in identifying more specific efficiency drivers.

5. **Operational Implementation of Dashboards**

   The Excel dashboard developed in this project can serve as a prototype for real-time business dashboards. Future versions could incorporate **live filters**, **automated updates** and executive-level KPI tracking for operational teams.

11. Bibliography

1. **Kaggle Dataset**. *Amazon delivery dataset* by Sujal Suthar.

   https://www.kaggle.com/datasets/sujalsuthar/amazon-delivery-dataset/data

2. **Microsoft Support**. *Used the Analysis ToolPak to perform complex data analysis*.

   https://support.microsoft.com/en-us/office/use-the-analysis-toolpak-to-perform-complex-data-analysis-6c67ccf0-f4a9-487c-8dec-bdb5a2cefab6

3. **Microsoft Support**. *Create a PivotTable to analyze worksheet data*.

   https://support.microsoft.com/en-us/office/create-a-pivottable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f99134456576

4. **Stylus.** *Calculate distance between two coordinates in Excel The Haversine formula*.

   https://www.stylus.co.za/calculate-the-distance-between-two-coordinates-in-excel/

5. **GeeksforGeeks**. *Data analysis in Excel: A beginner's guide*.

   https://www.geeksforgeeks.org/excel/data-analysis-in-excel/

6. **Simplilearn**. *Data analysis in Excel: The best guide*.

   https://www.simplilearn.com/tutorials/excel-tutorial/data-analysis-excel