**Bhakti Patil, Shashank Rewankar, Prithvi Bhatt, Arnav Laturkar**

**Project - Phase II: Decision Making**

**IFT 533: Data Visualization and Reporting for IT**

**Prof. Asmaa Elbadrawy**

**November 19, 2023**

**Section 1: Used Visualization Tools:**

To create dashboards, we use Tableau as our dashboard visualization tool. You may quickly view any item, chart, and dashboard in your database with this tool. Tableau is made to be easy to use and understand by people who are not technical or programmers. Its drag-and-drop elements simplify and rationalize every visualization.

Furthermore, Tableau has the ability to connect to a large variety of data sources that are not supported by other programs. Reports can be generated by users by integrating and merging various datasets. In addition, Tableau offers a number of choices, including the ability to split and concatenate fields, remove calculated fields, add headers and spaces, and modify data types. On the other hand, coding and a substantial amount of time would be needed to complete the same tasks in Python.

Tableau offers an infinite number of platforms for data collection and can connect to and extract data from many sources. This program offers multiple ways to extract data from spreadsheets, PDFs, Oracle databases, and cloud databases like Microsoft Azure SQL, Google Cloud SQL, and Amazon Web Services.

We particularly like Tableau's visual capabilities, which are one of its best features. Tableau offers visually appealing and useful visualizations that show data and assessments in an easily comprehensible manner. Its drag-and-drop feature facilitates the addition of dimensions, measurement of filters, grouping of data, exploration of outliers, and other tasks. Users may easily spot trends in their data with the use of advanced features like average lines, trend analysis, and clustering.

With Tableau's calculation builder, analysts can quickly and easily explore their data in novel ways. Visualization makes it easier to pinpoint problem areas and helps you get insights faster. A more thorough perspective of the data is provided by Tableau's extensive selection of chart styles, encodings, color-coding, and grouping options, all of which improve the dashboard experience overall.

**Section 2: Data-Preprocessing required**

1. Removing null values from the dataset for the attribute - brand.
2. Aggregating and summarizing the data for the attributes - model, year, color, fuel and transmission, kilometer and city.
3. Removing null values for attributes - kilometer and city.
4. Removing suffix in the attribute price and converting its data type to floating point value.

**Section 3: Final set of questions:**

1. Which are the most popular car models?
2. Which cities have the most customers that purchase cars? What are the most popular car models there? Which cities have the fewest customers?
3. What are the average distances traveled by cars that use different transmission modes and fuel types?
4. How does the average price of gasoline cars compare to diesel and hybrid cars?
5. How does the average price of the different car models compare to their respective mileages?
6. Which month of advertisement dates had the highest effect on the sales of cars?
7. Which car colors are the most popular among customers?
8. What are the most popular variants of the different car models?
9. What are the most popular colors for each car model?
10. What are the most popular colors for each variant of a car model?
11. What are the most popular transmission types in cars purchased? What are the counts of fuel types used by the car by different transmission types?
12. Which advertisement numbers brought in the highest number of car sales?
13. Which year of production of cars is the most popular among customers?
14. What are the distributions of years of production of different car models?
15. What are the average prices of the different variants of the same car models?
16. What are the averages prices of cars based on types of transmission mode used by the car?
17. Which transmission modes are the most popular in each city?

**Section 4: Dashboard Plot Drafts**

**1. Which are the most popular car models?**

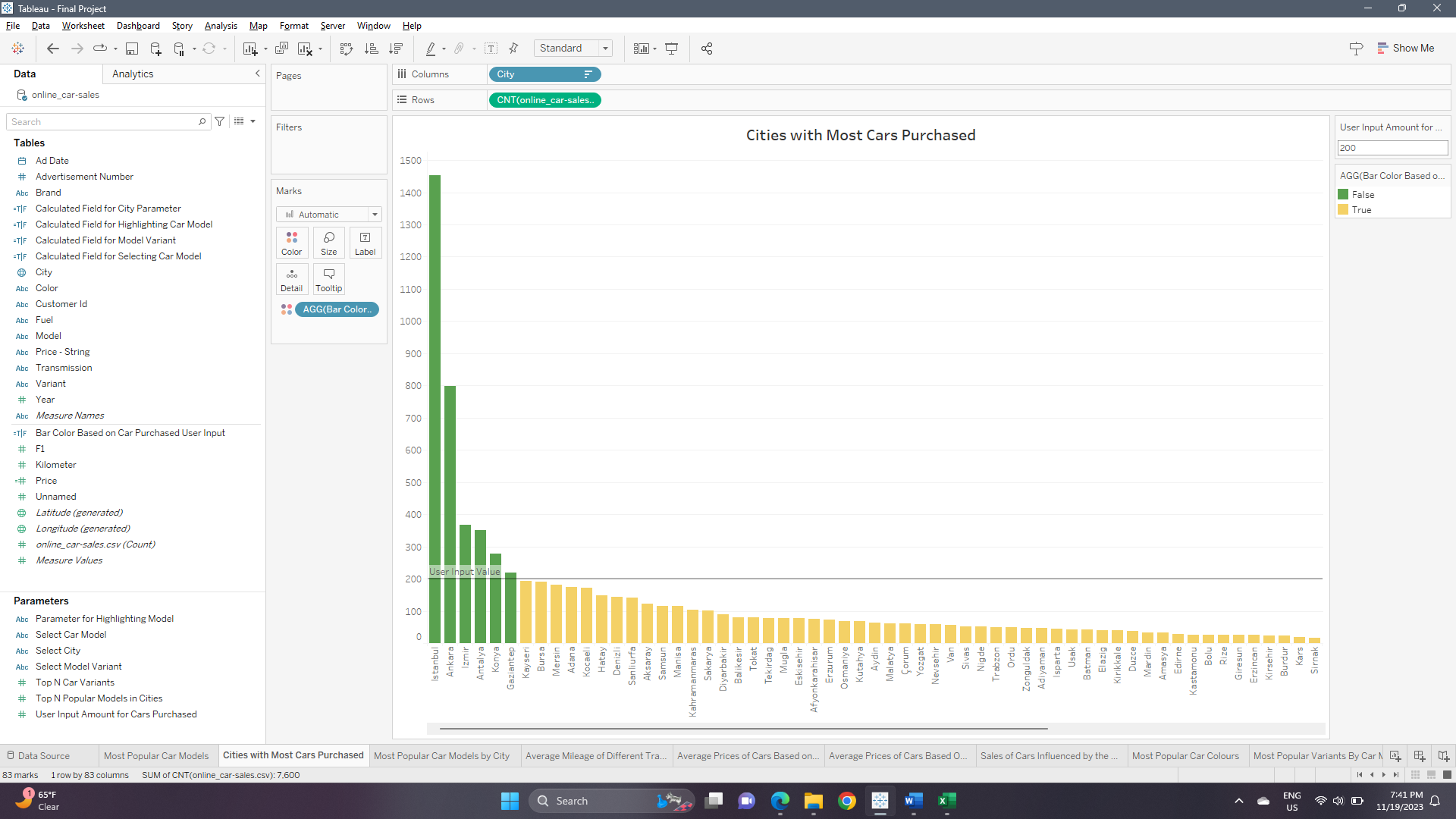
The plot used to address this question is a **bar chart** as it is the best choice to display the amounts of quantities belonging to different categories. The pre-attentive attribute used for this chart is **length**. Below is shown how the plot will look like:

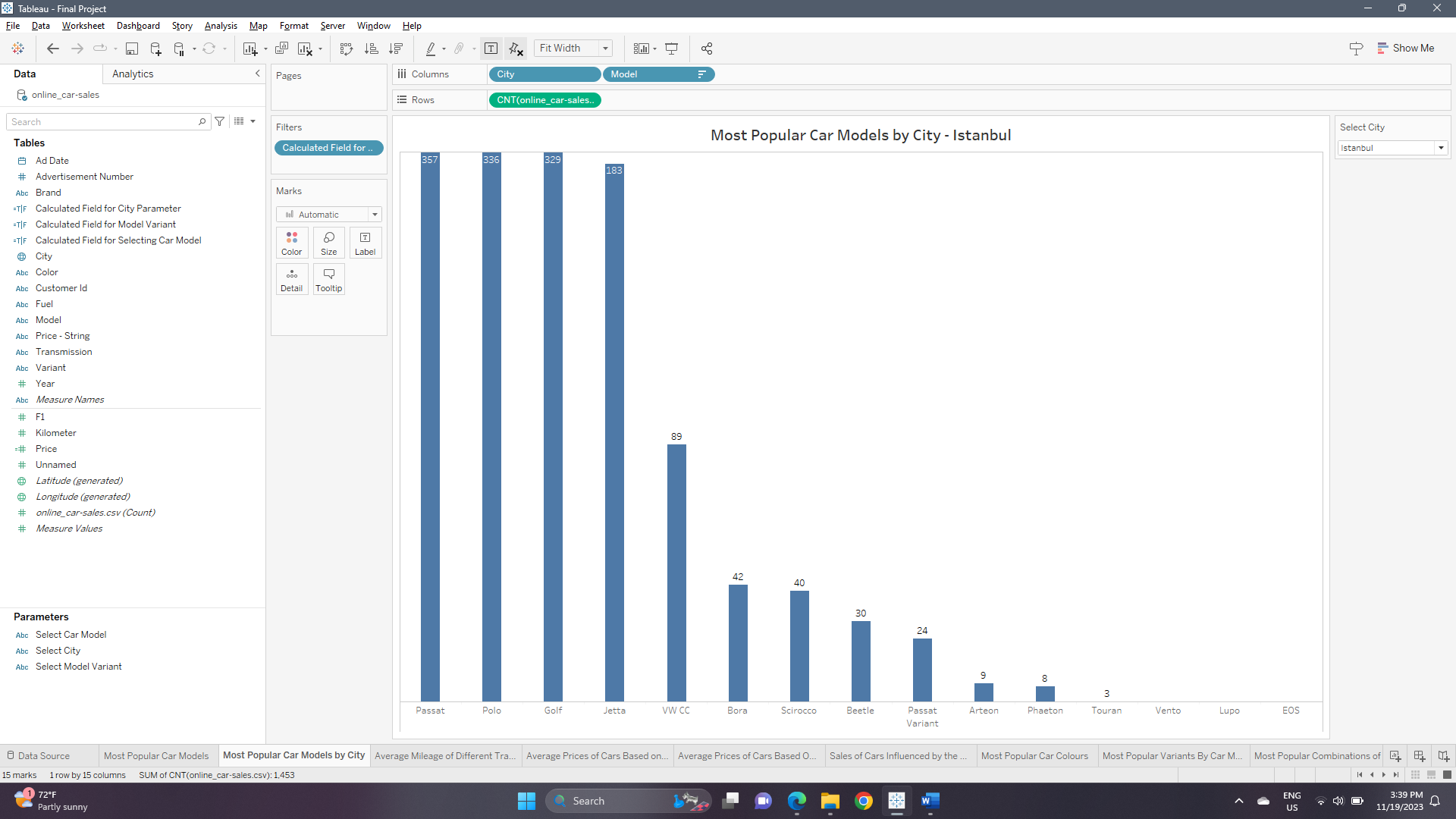
A screenshot of a computer

Description automatically generated

1. **Which cities have the most customers that purchase cars? What are the most popular car models there? Which cities have the fewest customers?**

The plots used to address both the questions are **bar charts** as it is the best choice to display the number of quantities belonging to different categories. We are also implementing a user-input parameter to specify an amount that will colour the bars based on the input provided. The pre-attentive attribute used for the first chart is **length**, and for the second chart is **length and color** (to highlight the cities that had customers purchasing more cars than specified user input). Below are shown how the plots will look like:





1. **What are the average distances traveled by cars that use different transmission modes and fuel types?**

The plot used to address this question is a **multi-line bar chart** as it is the best choice to display the amounts of quantities belonging to different categories of multiple attributes. The pre-attentive attribute used for this chart is **length**. Below is shown how the plot will look like:

A screenshot of a computer

Description automatically generated

1. **How does the average price of gasoline cars compare to diesel and hybrid cars?**

The plot used to address this question is a **scatter plot** as it is the best choice to compare the amounts of quantities belonging to different categories or numerical attributes. The pre-attentive attribute used for this chart is **position**. Below is shown how the plot will look like:

A screenshot of a computer

Description automatically generated

1. **How does the average price of the different car models compare to their respective mileages?**

The plot used to address this question is a **scatter plot** as it is the best choice to compare the amounts of quantities belonging to different categories or numerical attributes. The pre-attentive attributes used for this chart are **position and color** (for representing different car models). Below is shown how the plot will look like:

A screenshot of a computer

Description automatically generated

1. **Which month of advertisement dates had the highest effect on the sales of cars?**

The plot used to address this question is a **line chart** as it is the best choice to compare the amounts of quantities changing over a period of time. We have also added a reference line in the plot that denotes the average sales achieved over the whole time period. The pre-attentive attributes used for this chart are **position and orientation (of the line between consecutive points)**. Below is shown how the plot will look like:

A screenshot of a computer

Description automatically generated

1. **Which car colors are the most popular among customers?**

The plot used to address this question is a **bar chart** as it is the best choice to display the amounts of quantities belonging to different categories. The pre-attentive attribute used for this chart is **length**. Below is shown how the plot will look like:

A screenshot of a computer

Description automatically generated

1. **What are the most popular variants of the different car models?**

The plot used to address this question is a **bar chart** as it is the best choice to display the amounts of quantities belonging to different categories. We have added two parameters in this visualization; one for the user to select car model, and another is a ‘Top-N filter’ to give the user the option to specify how many top model variants should be displayed. The pre-attentive attribute used for this chart is **length**. Below is shown how the plot will look like:

A screenshot of a computer

Description automatically generated

1. **What are the most popular colors for each car model?**

The plot used to address this question is a **bar chart** as it is the best choice to display the amounts of quantities belonging to different categories. We have added one parameter in this visualization for the user to select car model for which the distribution of colours will be displayed. The pre-attentive attribute used for this chart is **length**. Below is shown how the plot will look like:

A screenshot of a computer

Description automatically generated

1. **What are the most popular colors for each variant of a car model?**

The plot used to address this question is a **bar chart** as it is the best choice to display the amounts of quantities belonging to different categories. We have added two parameters in this visualization; one for the user to select the car model, and another to select the model variant. The pre-attentive attribute used for this chart is **length**. Below is shown how the plot will look like:

A screenshot of a computer

Description automatically generated

1. **What are the most popular transmission types in cars purchased? What are the counts of fuel types used by the car by different transmission types?**

The plot used to address the first question is a **bar chart** as it is the best choice to display the amounts of quantities belonging to different categories. The pre-attentive attribute used for this chart is **length**. The plot used to address the second question is a **multi-line bar chart** as it is the best choice to display the amounts of quantities belonging to different categories of multiple attributes. The pre-attentive attribute used for both charts is **length**. Below are shown how both the plots will look like:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **Which advertisement numbers brought in the highest number of car sales?**

The plot used to address the question is a **bar chart** as it is the best choice to display the amounts of quantities belonging to different categories. The pre-attentive attribute used for this chart is **length**. Below are shown how both the plots will look like:

A screenshot of a computer

Description automatically generated

1. **Which year of production of cars is the most popular among customers?**

The plot used to address this question is a **line chart** as it is the best choice to compare the amounts of quantities changing over a period of time. The pre-attentive attributes used for this chart are **position and orientation (of the line between consecutive points)**. Below is shown how the plot will look like:

A screen shot of a computer

Description automatically generated

1. **What are the distributions of years of production of different car models?**

The plot used to address this question is a **multi-line or a line chart** as it is the best choice to compare the amounts of quantities changing over a period of time. We have added one parameter in this visualization, for the user to select car model for which the amounts of different years of production will be displayed. The plot will be a multi-line chart when the option ‘ALL’ is selected that displays the values for all models. If a single model is selected, it will be a single line chart. The pre-attentive attributes used for this chart are **position, orientation (of the line between consecutive points) and color** (to represent different car models). Below is shown how the plot will look like:

A screenshot of a computer

Description automatically generated

1. **What are the average prices of the different variants of the same car models?**

The plot used to address this question is a **scatter plot** as it is the best choice to compare the amounts of quantities belonging to different categories or numerical attributes. We have added one parameter in this visualization, for the user to select the car model. The pre-attentive attribute used for this chart is **position**. Below is shown how the plot will look like:

A screenshot of a computer

Description automatically generated

1. **What are the averages prices of cars based on types of transmission mode used by the car?**

The plot used to address the first question is a **bar chart** as it is the best choice to display the amounts of quantities belonging to different categories. The pre-attentive attribute used for this chart is **length**. Below are shown how both the plots will look like:

A screenshot of a computer

Description automatically generated

1. **Which transmission modes are the most popular in each city?**

The plot used to address the first question is a **bar chart** as it is the best choice to display the amounts of quantities belonging to different categories. We have added one parameter in this visualization, for the user to select the city for which the distribution of car transmission modes will be displayed. The pre-attentive attribute used for this chart is **length**. Below are shown how both the plots will look like:

A screenshot of a computer

Description automatically generated

**Section 5: Dashboard Interactivity**

We are implementing various interactive controls that will aid the user in better understanding the visualizations and are connected to various visualizations as required. Below is the list of interactive controls we have implemented –

* **Select Car Model –** This is a simple dropdown list that is used to specify the model of the car wherever required, and it is added in many visualizations. It is controlled using the calculated field ‘Calculated Field for Selecting Car Model’ that helps in filtering data based on the model selected. This parameter control is used in visualizations that answer the following questions – ‘What are the most popular variants of the different car models?’, ‘What are the most popular colors for each car model?’, ‘What are the most popular colors for each variant of a car model?’, ‘What are the distributions of years of production of different car models?’, and ‘What are the average prices of the different variants of the same car models?’. The values for this parameter are loaded from the ‘Model’ attribute of the dataset. It is a string attribute and hence has no range values.
* **Select Model Variant –** This is a simple dropdown list that is used to specify the variant of a particular car model wherever required, and it is added in one visualization. It is controlled using the calculated field ‘Calculated Field for Selecting Model Variant’ that helps in filtering data based on the variant selected. This parameter control is used in only one visualization that answers the following questions – ‘What are the most popular colors for each variant of a car model?’. The values for this parameter are loaded from the ‘Variant’ attribute of the dataset. It is a string attribute and hence has no range values. Another important property of this parameter is that it has been configured to display only relevant values based on the car model that has been selected using the ‘Context’ feature provided in Tableau. This helps in displaying only the variants that are particular to a specific car model.
* **Select City –** This is a simple dropdown list that is used to specify the city where the car was purchased wherever required, and it is added in two visualizations. It is controlled using the calculated field ‘Calculated Field for City Parameter’ that helps in filtering data based on the city selected. This parameter control is used in visualizations that answer the following questions – ‘Which cities have the most customers that purchase cars? What are the most popular car models there? Which cities have the fewest customers?’ and ‘Which transmission modes are the most popular in each city?’. The values for this parameter are loaded from the ‘City’ attribute of the dataset. It is a string attribute and hence has no range values.
* **Top-N Car Variants –** This is a slider provided for the user to specify the number of most popular variants of a particular car model that should be displayed as required. This feature is added to one visualization that answers the question – ‘What are the most popular variants of the different car models?’. This parameter only allows Integer values, with a range from a minimum of ‘1’ to a maximum of ‘10’ with a step-size of ‘1’.
* **Top-N Popular Models in Cities –** This is a slider provided for the user to specify the number of most popular car models that should be displayed as required. This feature is added to one visualization that answers the question – ‘Which cities have the most customers that purchase cars? What are the most popular car models there?’. This parameter only allows Integer values, with a range from a minimum of ‘1’ to a maximum of ‘10’ with a step-size of ‘1’.
* **User Input Amount for Cars Purchased –** This is a ‘Type-In’ input parameter created to allow the user to specify a particular number of cars purchased in a city. This parameter control is added to the visualization that answers the question – ‘Which cities have the most customers that purchase cars?’. The value provided by the user is compared with each value for a city that is used to create a bar, and then a reference line is added at that particular value. The bars of cities that have a higher number of cars purchased are then colored green, and the remaining bars are colored gold. A calculated field ‘Bar Color Based on Cars Purchased User Input’ is created and added as a filter to the visualization to make it work as intended. This parameter only allows Integer values, with a range from a minimum of ‘10’ to a maximum of ‘1500’ with a step-size of ‘10’.

**Section 6: References**

**Link to Dataset –**

<https://www.kaggle.com/datasets/bimervos/online-automotive-sales-statistics-volkswagen>

**Link to Mural –**

<https://app.mural.co/t/dvassignment0284/m/dvassignment0284/1699563981327/42f506657e2cd9915e45641bf72383a216df88c2?sender=ua05eef829bc2a8f507a20491>