Homework Assignment 1

Data for Supporting BI

Part 1) Essay: In 1-2 pages explain how and why the seven data elements discussed in the reading and in the lecture are important and should be handled properly in the context of the company (or non-profit organization) that you focused on in the online discussion last week (data-driven innovations). For example, if you focused on Starbucks, now explain the implications of data structure (Do you think Starbucks uses, or should use, both structured and unstructured data in their business intelligence and analytics processes?), data uniqueness (Can Starbucks’ data or data analytics capabilities provide a source of competitive advantage for them? Why and how?), data integration (Why is it important for them to consolidate data from different sources? What would be a challenge for them to do so?), data quality (What data quality issues do you think they may need to handle and why?), data access (Why is it important for Starbucks to ensure data availability from an analytics perspective?), data privacy (What are the implications of data privacy in the context of Starbucks, particularly when it comes to analytics), data governance (What does data governance mean in the context of Starbucks and why does it matter?) Your explanations and arguments should be relevant in the context of the company that you worked on in the online discussion last week (Starbucks was just an example). The explanations should be written in the form of an essay and submitted as part of this assignment. The essay should start with an introduction (a short paragraph or few sentences) describing and introducing the context that you are focusing on (e.g., if it is Allstate Insurance Company, briefly explain what that company does and where they are located at).

Delta Air Lines, a major player in the aviation industry, serves as the focal point for this essay. Headquartered at 1030 Delta Blvd, Atlanta, Georgia, Delta operates a vast global network, catering to nearly 200 million passengers annually. As we explore the importance of seven key data elements - data structure, data uniqueness, data integration, data quality, data access, data privacy, and data governance - within the context of Delta Airlines, we aim to understand how handling these elements appropriately is crucial for fostering data-driven innovations in the aviation sector.

1. Data Structure:

Within Delta Airlines, effective data analytics relies on the utilization of both structured and unstructured data. Structured data, encompassing information like flight schedules, crew details, and passenger records, forms the backbone for traditional analytics. Simultaneously, unstructured data, derived from sources such as social media, customer feedback, and weather reports, contributes to a more comprehensive understanding of operational challenges and customer preferences. A strategic integration of structured and unstructured data ensures that Delta can derive actionable insights for enhancing operational efficiency and customer experience.

1. Data Uniqueness:

Delta's data uniqueness, shaped by its extensive operational data, customer interactions, and industry insights, becomes a valuable source of competitive advantage. The airline's ability to harness unique data sets allows for personalized customer services, optimized flight schedules, and strategic decision-making. The insights derived from Delta's distinct data contribute to innovations in route planning, customer engagement, and service delivery, reinforcing the company's competitive position in the aviation industry.

1. Data Integration:

Given the complexity of the aviation industry, Delta Airlines benefits significantly from integrating data from various sources. Data related to flight operations, crew scheduling, weather conditions, and customer preferences needs to be consolidated for a holistic view. Challenges arise due to the diverse formats and structures of data received from different sources. Streamlining data integration processes ensures that Delta can make informed decisions, optimize resource allocation, and enhance overall operational efficiency.

1. Data Quality:

High data quality is paramount for Delta to ensure the accuracy and reliability of its analytics processes. In the aviation sector, inaccuracies or incomplete data can lead to severe consequences. Delta must address data quality issues such as discrepancies in flight schedules, crew information, and maintenance records. By implementing robust data quality measures, Delta enhances the precision of its analytics outputs, resulting in improved safety, on-time performance, and customer satisfaction.

1. Data Access:

Timely access to accurate and relevant data is critical for Delta Airlines. Operational decisions, crew management, and customer interactions benefit from real-time analytics. Establishing efficient data access mechanisms allows Delta to respond swiftly to dynamic operational challenges, optimize flight operations, and deliver an enhanced travel experience. Ensuring data availability supports Delta in making agile, data-driven decisions in a fast-paced industry.

1. Data Privacy:

Delta Airlines, dealing with sensitive passenger information and operational data, must prioritize stringent data privacy measures. Compliance with privacy regulations is crucial to maintain customer trust and adhere to legal requirements. Ensuring that analytics processes uphold data privacy safeguards Delta from potential risks, builds customer confidence, and strengthens the airline's commitment to ethical data handling.

1. Data Governance:

In the context of Delta, data governance involves establishing comprehensive policies and procedures to ensure responsible data management. This encompasses defining data ownership, setting quality standards, and ensuring compliance with industry regulations. Effective data governance supports Delta in maintaining data integrity, facilitating collaboration across departments, and upholding ethical data practices throughout the organization.

In conclusion, the appropriate handling of data structure, data uniqueness, data integration, data quality, data access, data privacy, and data governance is indispensable for Delta Airlines. By prioritizing these elements, Delta can harness the power of data-driven innovations, optimize operational processes, and continue providing exceptional services in the competitive aviation landscape.

1. (Excel) How many customers do we have in the table?

The dataset contains a comprehensive record of 1,180 customers associated with a consumer product company, as identified by the unique CustomerID information.

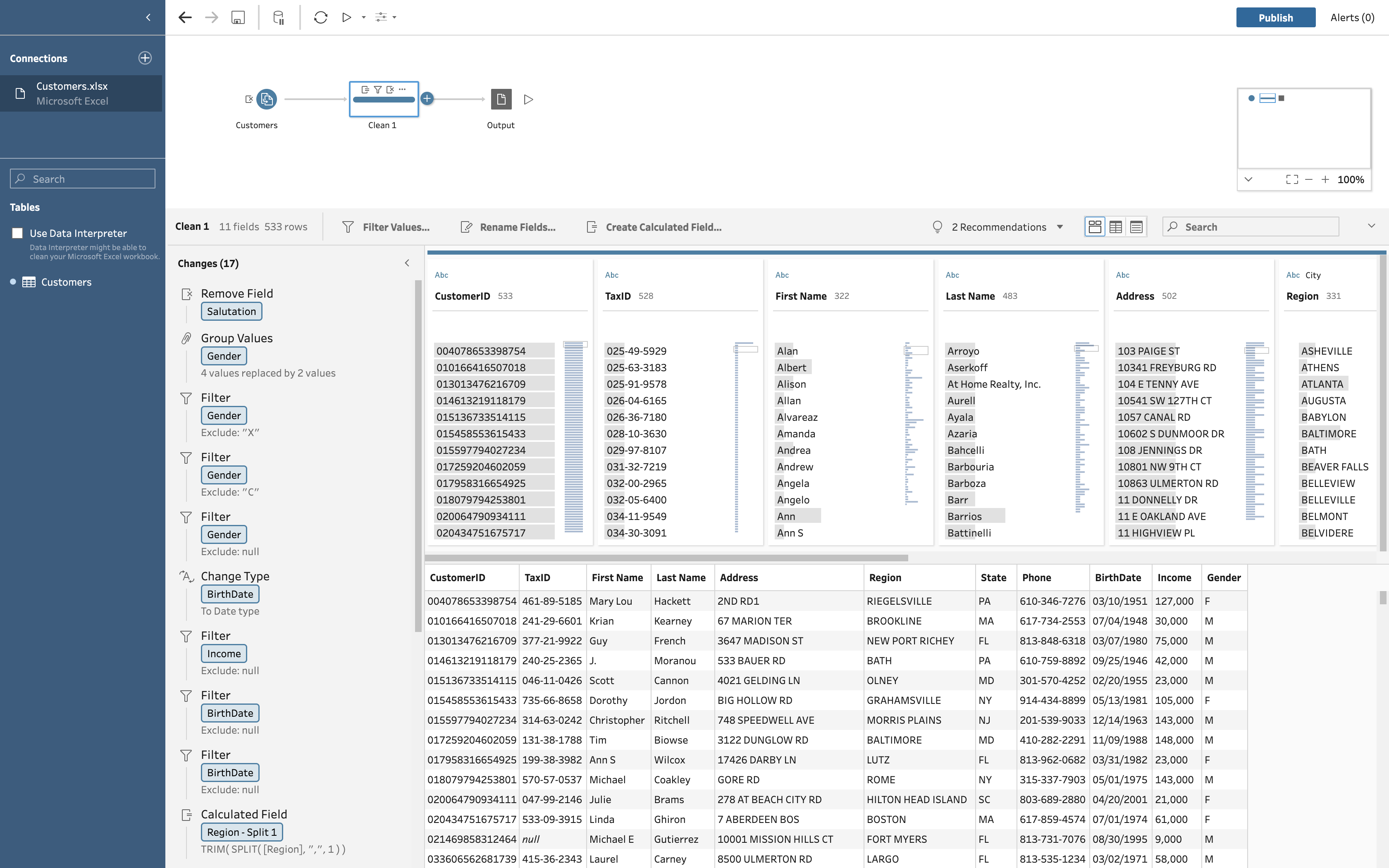
1. (Excel) What are the quality issues with the customer data? Explain how you identified each issue.

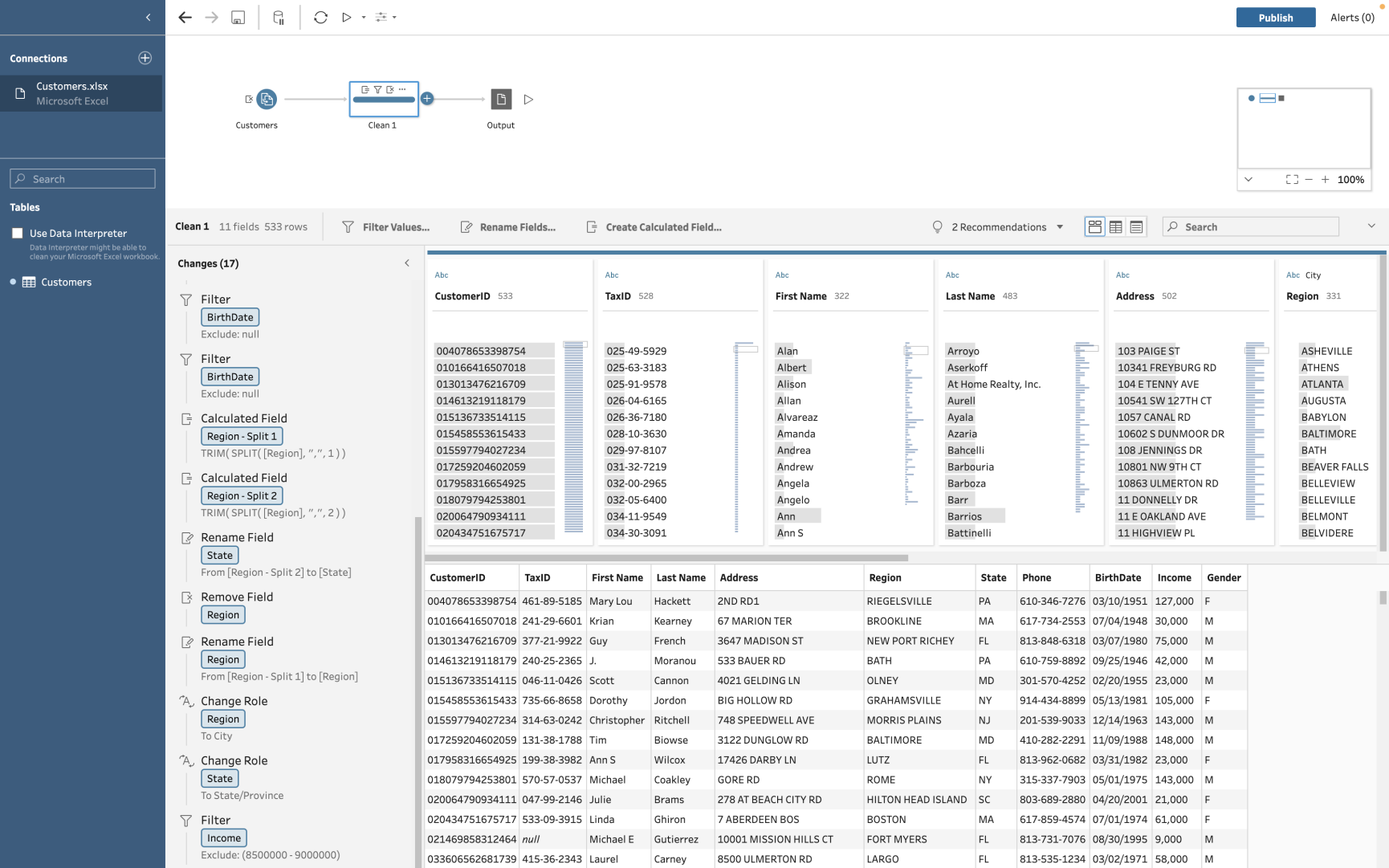
Upon detailed examination of the dataset, the following quality issues with the customer data were identified and addressed:

* Duplicate Records: No instances of duplicate records were found in the dataset, ensuring data integrity.
* TaxID: The TaxID field, unique for each customer, displayed 17 blank records, constituting approximately 1% of the dataset.
* Salutation: The Salutation field was found to be completely blank and was deemed irrelevant for analysis.
* First Name: The First Name field, containing sensitive customer information, exhibited 42 blank records, equivalent to approximately 4% of the dataset.
* Last Name, Address, and Region: These fields, containing sensitive customer information, were properly populated. However, the Region column was noted to consolidate information about both region and state, suggesting a potential need for further data refinement.
* Phone: The Phone column, containing customer-sensitive information, displayed consistent 10-character phone number details for each customer.
* BirthDate: The BirthDate field, which includes sensitive information, revealed 74 blank records, making up approximately 6% of the dataset. Furthermore, the column contained negative numbers and date values in corrupted formats, indicating a need for data cleansing.
* Income: The Income field, also sensitive, presented 76 blank records, constituting around 6% of the dataset. Additionally, an outlier value of "8881000" was identified and recommended for removal after percentile distribution analysis.
* Gender: The Gender field, containing sensitive information, had 492 blank records, representing approximately 42% of the dataset. During data cleaning, non-sensical values such as "C" and "X" were removed, and variations in gender terms were standardized (e.g., "Femal", "Female", "Fmale" to "F" and "Male" to "M").

In conclusion, the comprehensive analysis and data cleansing efforts undertaken have significantly enhanced the quality and reliability of the customer data, addressing various issues to ensure its suitability for further analysis and decision-making.

1. (Tableau Prep) Using Tableau Prep Builder clean the dataset and generate an Excel file as an output. Explain each data cleaning step and provide a final screenshot of the Tableau Prep Builder interface with the flow chart and profile pane visible on it to help us understand how you have done it. As part of the data preparation process, split the column Region into two columns (City and State) as you will need to use this information to answer some of the questions below. Change the new columns’ data roles to City and State/Province. To learn more about data roles see here. Next, perform any grouping as required to correct data entry errors within the fields (e.g., gender). For gender, group together values that likely represent the same gender (e.g., Female and F, but not F and X) based on your judgment. Finally, check each field to ensure correct data roles and types (City and State should be geographic, Birthdate should be Date, etc.).





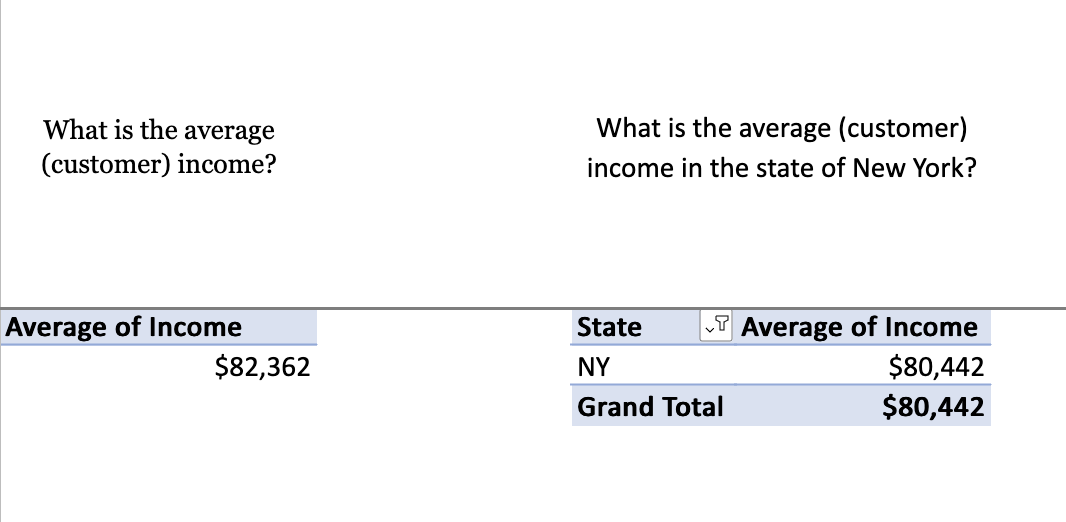
1. (Excel) Based on the cleaned dataset and using functions, pivot tables, and charts in Excel answer the following questions (for each item, even the items that require a short answer such as item a, provide a brief explanation and if applicable, a screenshot of your Excel output (e.g., as part of a table/pivot table, a chart, or a formula) to show how you have done/calculated it).

a) What is the average (customer) income?

The average income for customers is **$82,362**.

b) What is the average (customer) income in the state of New York?

The average income for customers in the state of New York is **$80,442**.



c) Which three states have the highest average incomes? Which three states have the lowest average incomes?

**Highest Average Incomes:**

- Alabama (AL) - $125,250

- West Virginia (WV) - $101,333

- Tennessee (TN) - $101,167

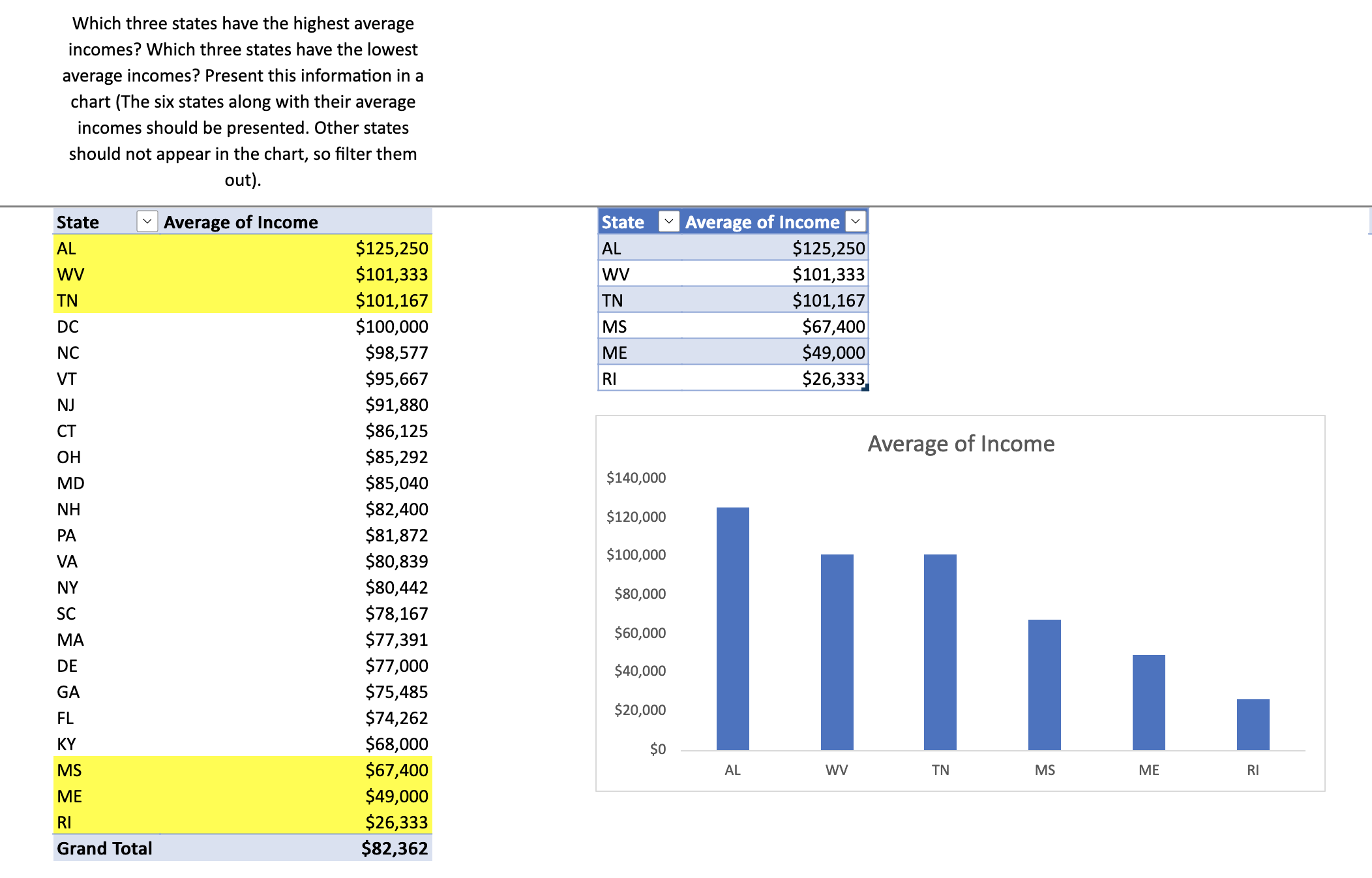
**Lowest Average Incomes:**

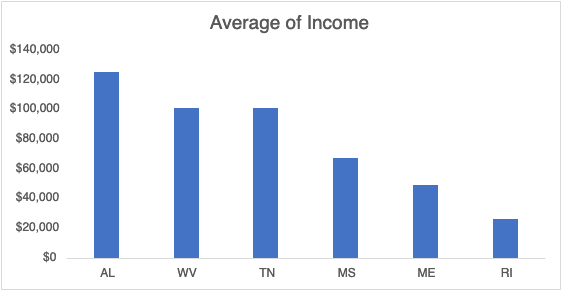
- Mississippi (MS) - $67,400

- Maine (ME) - $49,000

- Rhode Island (RI) - $26,333

|  |  |  |
| --- | --- | --- |
| Sr. No | State | Average Income |
| 1 | AL | $125,250 |
| 2 | WV | $101,333 |
| 3 | TN | $101,167 |
| 4 | MS | $67,400 |
| 5 | ME | $49,000 |
| 6 | RI | $26,333 |



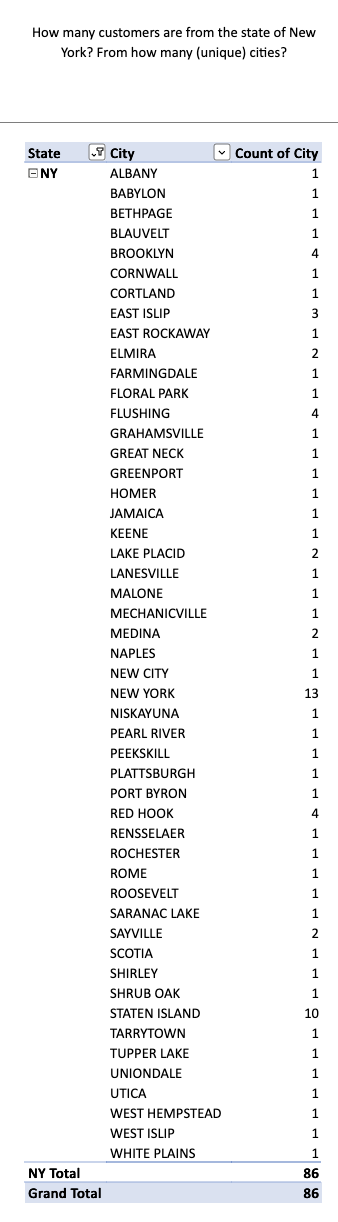


d) Did you remove any outliers before calculating the average incomes (parts a-c)? Explain.

Yes, before calculating the average incomes, the outlier value of "**8881000**" in the dataset was identified through percentile distribution analysis and subsequently removed.

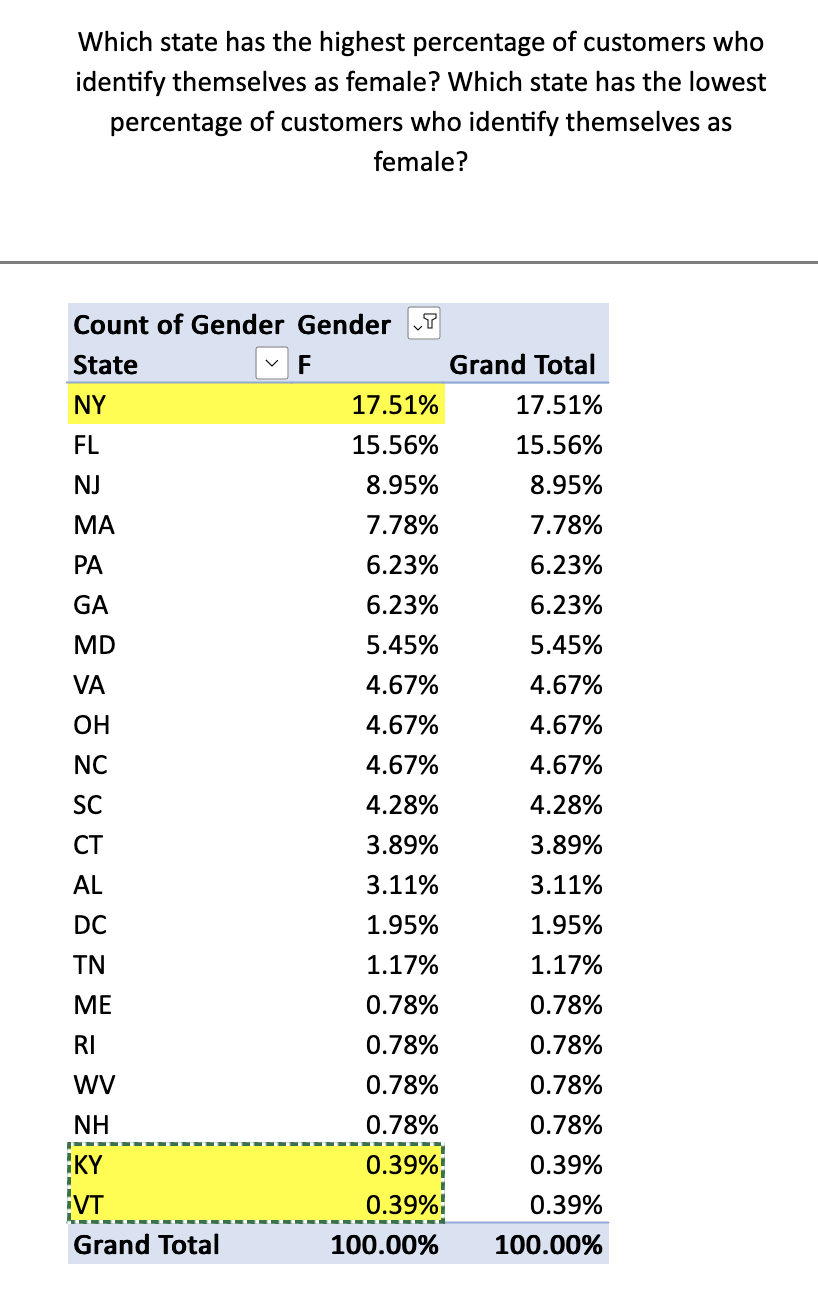
e) How many customers are from the state of New York? From how many (unique) cities?

There are **86** customers from the state of New York, distributed across a total of **50** unique cities.



f) Which state has the highest percentage of customers who identify themselves as female? Which state has the lowest percentage of customers who identify themselves as female?

New York (**NY**) boasts the highest percentage of customers who identify themselves as female, with **17.51%**. On the other hand, Kentucky (**KY**) and Vermont (**VT**) both have the lowest percentage, with **0.39%** each.



1. (Excel or Tableau Prep) Did you perform any listwise or pairwise deletion during the data cleaning process? How would performing each of these techniques impact your answers to part 4 (items a – f)? (You do not need to perform both; whichever you have done, explain how using the other one would change the results.) Would it be possible and useful to run any imputation on the dataset? (Again, you do not have to actually do it; just explain.)

In the data cleaning phase, a combination of listwise deletion and pairwise deletion methodologies was employed to address missing values. Listwise deletion involved the removal of entire rows with missing values, while pairwise deletion considered available data for specific analyses.

Given the prevalence of blank or null values in most columns, a decision was made to remove such data before individual column processing. Although data removal is typically considered a last resort, the nature and scale of the dataset warranted this approach before in-depth column-wise analysis. Consequently, this resulted in a reduced sample size and varying statistical values for each column during subsequent analyses.

a) Average Income:

- Listwise Deletion: Rows with missing income values were excluded, potentially yielding a diminished sample size and altered average income.

- Pairwise Deletion: Only available income values were considered post-data removal, leading to fluctuating averages based on data completeness.

b) Average Income in New York:

- A similar impact to the above example would be observed.

c) Average Incomes by State (Highest and Lowest):

- The listwise or pairwise deletion approach may affect states' average incomes if missing values are not uniformly distributed across states, introducing potential bias.

d) Outlier Removal:

- The outlier removal process remains unaffected by listwise or pairwise deletion, as outliers are specific values rather than missing values.

e) Number of Customers from New York:

- The count might be influenced if listwise deletion was applied, removing entire rows with missing state information.

f) Percentage of Female Customers by State:

- The percentage may be impacted if listwise deletion disproportionately eliminates rows with missing gender information, affecting the overall distribution.

Imputation:

- Given the dataset's nature and the sensitivity of customer information, imputation may not be advisable for certain fields such as income or birthdate.

In conclusion, the selection between listwise and pairwise deletion hinges on the distribution of missing values and the potential introduction of bias. Imputation, while applicable to less sensitive fields, necessitates a cautious approach, taking into account data nature and its implications on privacy and accuracy.