# A PROJECT TO STUDY AND GENERATE A CUSTOMER 360° VIEW STORY USING TABLEAU

Project submitted and prepared by Bishal Biswas
under guidance of Prof. Sambita Chakraborty

Project done as a part of assignments for
Post Graduate Diploma Program
From
Bombay Stock Exchange (BSE)
In collaboration with
Maulana Abul Kalam Azad University of Technology
(MAKAUT)

# Certification of Approval

This document is hereby approved as credible study of the science subject carried out and represented in a manner to satisfy to the warrants of its acceptance as a prerequisite to the degree for which it has been submitted.

Moreover, it is understood that by this approval the undersigned does not necessarily endorse or approve any statements made, the opinion expressed or conclusion drawn therein but approved only for the sole purpose for which it has been indeed submitted.

Signatures of the Examiners with date.

×_____

×_____

×_____

Dated:

Countersigned by:

×_____

Prof. Sambita Chakraborty

# Acknowledgement

Our project and everything started during the ongoing reign of SARS – CoVID19, virtually crippling the society and world as a whole sending everything into a lockdown. Although at better stage but with a second wave looming on the pre-existing distress in-spite of new vaccines combating the situation. Overcoming all this chaos the course of Post Graduate Diploma in Data Science by BSE in collaboration with MAKAUT was made possible thanks to the diplomacy and steps taken by both institutes to combat the situation and make this course and project a possibility.

I want to take this opportunity of the project to thank the people at BSE and MAKAUT who provided us this opportunity to have an exposure to real life scenarios and the status of the present market. I also want to thank Prof. Sambita Chakraborty for guiding with every step from imparting knowledge about the subject to the intricacies of the Data Preparation and Analysis including Cleaning, Visualization, clearing doubts and issues faced in addition to solving problems encountered.

I am also grateful to my batchmates and peers where our collective knowledgebase and doubt clearing helped a lot in completing this project. Lastly, I want to thank my family for the mental support they provided me that played a big part in completing this project.

×_____

Bishal Biswas.
BSE GENERATED ID: PGDDSPJULY2020/1
MAKAUT ENROLMENT:20BIL001P12029005
MAKAUT APPLICATION ID: 91268
b.biswas_94587@ieee.org

# Contents


| | |
|---|---|
| Objective and Purpose | 4 |
| Introduction | 5 |
| Data Gathering | 7 |
| Data Preparation | 8 |
| Data Cleaning | 10 |
| Data Analysis | 11 |
| Exploratory Data Analysis ~ EDA | 13 |
| Data Interpretation | 15 |
| Data Visualization | 18 |
| A little intro the problem at hand | 27 |
| TABLEAU | 28 |
| Customer 360° view | 29 |
| Problem Statement | 30 |
| General Idea of the Data Received | 30 |
| Problems in the Data | 31 |
| The Correcting Code | 38 |
| A look into the presentable 360-degree analysis | 39 |
| Chart lists | 46 |
| Critical understanding achieved from the project | 48 |
| Bibliography | 49 |

## Objective and Purpose

Data is truly considered a resource in today's world. Data is everywhere and part of our daily lives in more ways than most of us realize in our daily lives. The amount of digital data that exists—that we create—is growing exponentially. As per the World Economic Forum, by 2025 we will be generating about 463 exabytes of data globally per day. Hence, there is a need for professionals who understand the basics of data science, big data, and data analytics. These three terms are often heard frequently in the industry, and while their meanings share some similarities, they also mean different things.

Data science is the combination of statistics, mathematics, programming, problem-solving, capturing data in ingenious ways, the ability to look at things differently, and the activity of cleansing, preparing, and aligning data. This umbrella term includes various techniques that are used when extracting insights and information from data.

Now, Big data refers to significant volumes of data that cannot be processed effectively with the traditional applications that are currently used. The processing of big data begins with raw data that isn't aggregated and is most often impossible to store in the memory of a single computer. A buzzword that is used to describe immense volumes of data, both unstructured and structured, big data can inundate a business on a day-to-day basis. Big data is used to analyze insights, which can lead to better decisions and strategic business moves.

Gartner provides the following definition of big data: "Big data is high-volume, and high-velocity or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

Data analytics involves applying an algorithmic or mechanical process to derive insights and running through several data sets to look for meaningful correlations. It is used in several industries, which enables organizations and data analytics companies to make more informed decisions, as well as verify and disprove existing theories or models. The focus of data analytics lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows.

Industries like IT, Retail, Manufacturing, Automobile, Financial Institute, E Commerce etc. are focusing in depth towards Big Data Concept because they have found out its importance, they know Data is Asset and its value will grow day by day and it can lead the Global business. Some benefits of it are:

- Data driven decision making with more accuracy.
- Customer active engagement.
- Operation optimization.

- Data driven Promotions.
- Preventing frauds & threats.
- Exploring new sources of revenue.
- Being ahead of your competitors.

Thus, the growth of big data analytics will also probably be good for data scientists, especially those who have strong backgrounds in big data. Based on the growth of the big data analytics market in the past few years, along with the rising number of job openings, it's likely that demand for these skills will continue to increase in the near future.

## Introduction

Since the invention of computers, people have used the term data to refer to computer information, and this information was either transmitted or stored. But that is not the only data definition; there exist other types of data as well. So, what is the data? Data can be texts or numbers written on papers, or it can be bytes and bits inside the memory of electronic devices, or it could be facts that are stored inside a person's mind.

Now, if we talk about data mainly in the field of science, then the answer to "what is data" will be that data is different types of information that usually is formatted in a particular manner. All the software is divided into two major categories, and those are programs and data. Programs are the collection made of instructions that are used to manipulate data. So, now after thoroughly understanding what is data and data science, let us learn some fantastic facts.

Growth in the field of technology, specifically in smartphones has led to text, video, and audio is included under data plus the web and log activity records as well. Most of this data is unstructured.

The term Big Data is used in the data definition to describe the data that is in the petabyte range or higher. Big Data is also described as 5Vs: variety, volume, value, veracity, and velocity. Nowadays, web-based eCommerce has spread vastly, business models based on Big Data have evolved, and they treat data as an asset itself. And there are many benefits of Big Data as well, such as reduced costs, enhanced efficiency, enhanced sales, etc.

The meaning of data expands beyond the processing of data in computing applications. When it comes to what data science is, a body made of facts is called data science. Accordingly, finance, demographics, health, and marketing also have different meanings of data, which ultimately make up different answers for what is data.

When we talk about data, we usually think of some large datasets with huge number of rows and columns. While that is a likely scenario, it is not always the case — data could be in so many different forms: Structured Tables, Images, Audio files, Videos etc. Machines don't understand free text, image or video data as it is, they understand 1s and 0s. So, it probably won't be good enough if we put on a slideshow of all our images and expect our machine learning model to get trained just by that, hence a need for processing the data is required before being fed to the system for any analysis or estimation making learning or prediction too farfetched an idea to see realization.

Clean data is crucial for insightful data analysis. Data cleansing, data cleaning or data scrubbing is the first step in the overall data preparation process. It is the process of analyzing, identifying and correcting messy, raw data. Data cleaning involves filling in missing values, identifying and fixing errors and determining if all the information is in the right rows and columns. When analyzing organizational data to make strategic decisions you must start with a thorough data cleansing process. Cleaning data is crucial to data analysis. Data cleaning lays the groundwork for efficient, accurate and effective data analysis. Without cleaning data beforehand, the analysis process won't be clear or as accurate because the information in the dataset will be unorganized and scattered. Good analysis rests on clean data–it's as simple as that.

Analysis is the process of breaking a complex topic or substance into smaller parts in order to gain a better understanding of it. The technique has been applied in the study of mathematics and logic since before Aristotle, though analysis as a formal concept is a relatively recent development. Implementing these ideas of analysis using statistical processes to determine the future or optimize situations using available data at the disposal or data obtained from a specific source is the very idea on which the Data Analysis and subsequently Big Data Analysis is based on.

Beyond the analysis, even if we stop at this point, we are just holding a set of numbers, some arranged others not but all equivalently shared between fields of significance, which if not properly visualized wouldn't make any sense to a person working beyond the realm or field from where the data is based. Hence, the need for visualization stands. Visualization is the process of putting together visual mental imagery of what you are wanting to manifest. Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

# Data Gathering

Data gathering or Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes.

Data collection can be carried by from various ways like:

1. Surveys. Surveys are one way in which you can directly ask customers for information.
2. Online Tracking.
3. Transactional Data Tracking.
4. Online Marketing Analytics.
5. Social Media Monitoring.
6. Collecting Subscription and Registration Data.
7. In-Store Traffic Monitoring.

## Primary Data Collection

The term "primary data" refers to data you collect yourself, rather than data you gather after another party initially recorded it. Primary data is information obtained directly from the source. You will be the first party to use this exact set of data.

When it comes to data businesses collect about their customers, primary data is also typically first-party data. First-party data is the information you gather directly from your audience. It could include data you gathered from online properties, data in your customer relationship management system or non-online data you collect from your customers through surveys and various other sources.

First-party data differs from second-party and third-party data. Second-party data is the first-party data of another company. You can purchase second-party data directly from the organization that collected it or buy it in a private marketplace. Third-party data is information a company has pulled together from numerous sources. You can buy and sell this kind of data on a data exchange, and it typically contains a large number of data points. Because first-party data comes directly from your audience, you can have high confidence in its accuracy, as well as its relevance to your business.

Second-party data has many of the same positive attributes as first-party data. It comes directly from the source, so you can be confident in its accuracy, but it also gives you insights you couldn't get with your first-party data. Third-party data offers much more scale than any other type of data, which is its primary benefit.

Different types of data can be useful in different scenarios. It can also be helpful to use different types of data together. First-party data will typically be the

foundation of your dataset. If your first-party data is limited, though, you may want to supplement it with second-party or third-party data. Adding these other types of data can increase the scale of your audience or help you reach new audiences.

### Quantitative vs. Qualitative Data

Quantitative data comes in the form of numbers, quantities and values. It describes things in concrete and easily measurable terms. Examples include the number of customers who bought a given product, the rating a customer gave a product out of five stars and the amount of time a visitor spent on your website.

Because quantitative data is numeric and measurable, it lends itself well to analytics. When you analyze quantitative data, you may uncover insights that can help you better understand your audience. Because this kind of data deals with numbers, it is very objective and has a reputation for reliability.

Qualitative data is descriptive, rather than numeric. It is less concrete and less easily measurable than quantitative data. This data may contain descriptive phrases and opinions. Examples include an online review a customer writes about a product, an answer to an open-ended survey question about what type of videos a customer likes to watch online and the conversation a customer had with a customer service representative.

Qualitative data helps explains the "why" behind the information quantitative data reveals. For this reason, it is useful for supplementing quantitative data, which will form the foundation of your data strategy. Because quantitative data is so foundational, this article will focus on collection methods for quantitative primary data.

## Data Preparation

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data. Good data preparation allows for efficient analysis, limits errors and inaccuracies that can occur to data during processing, and makes all processed data more accessible to users. It's also gotten easier with new tools that enable any user to cleanse and qualify data on their own.

Data preparation is often a lengthy undertaking for data professionals or business users, but it is essential as a prerequisite to put data in context in order to turn it into insights and eliminate bias resulting from poor data quality.

## Benefits of Data Preparation

76% of data scientists say that data preparation is the worst part of their job, but the efficient, accurate business decisions can only be made with clean data. Data preparation helps:

- Fix errors quickly — Data preparation helps catch errors before processing. After data has been removed from its original source, these errors become more difficult to understand and correct.
- Produce top-quality data — Cleaning and reformatting datasets ensures that all data used in analysis will be high quality.
- Make better business decisions — Higher quality data that can be processed and analyzed more quickly and efficiently leads to more timely, efficient and high-quality business decisions.

## Integrating the Cloud technology to the data preparation

As data and data processes move to the cloud, data preparation moves with it for even greater benefits, such as:

- Superior scalability — Cloud data preparation can grow at the pace of the business. Enterprise don't have to worry about the underlying infrastructure or try to anticipate their evolutions.
- Future proof — Cloud data preparation upgrades automatically so that new capabilities or problem fixes can be turned on as soon as they are released. This allows organizations to stay ahead of the innovation curve without delays and added costs.
- Accelerated data usage and collaboration — Doing data prep in the cloud means it is always on, doesn't require any technical installation, and lets teams collaborate on the work for faster results.

Considering beyond the above points, a good, cloud-native data preparation tool will offer other benefits (like an intuitive and simple to use GUI) for easier and more efficient preparation.

## Data Preparation Steps

The specifics of the data preparation process vary by industry, organization and need, but the framework remains largely the same.

1. Gather data: The data preparation process begins with finding the right data. This can come from an existing data catalog or can be added ad-hoc.

2. Discover and assess data: After collecting the data, it is important to discover each dataset. This step is about getting to know the data and understanding what has to be done before the data becomes useful in a particular context.
3. Cleanse and validate data: Cleaning up the data is traditionally the most time-consuming part of the data preparation process, but it's crucial for removing faulty data and filling in gaps. Important tasks here include:
    a. Removing extraneous data and outliers.
    b. Filling in missing values.
    c. Conforming data to a standardized pattern.
    d. Masking private or sensitive data entries.

    Once data has been cleansed, it must be validated by testing for errors in the data preparation process up to this point. Often times, an error in the system will become apparent during this step and will need to be resolved before moving forward.
4. Transform and enrich data: Transforming data is the process of updating the format or value entries in order to reach a well-defined outcome, or to make the data more easily understood by a wider audience. Enriching data refers to adding and connecting data with other related information to provide deeper insights.
5. Store data: Once prepared, the data can be stored or channeled into a third-party application—such as a business intelligence tool—clearing the way for processing and analysis to take place.

## Data Cleaning

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This data is usually not necessary or helpful when it comes to analyzing data because it may hinder the process or provide inaccurate results. There are several methods for cleaning data depending on how it is stored along with the answers being sought.

Data cleaning is not simply about erasing information to make space for new data, but rather finding a way to maximize a data set's accuracy without necessarily deleting information. For one, data cleaning includes more actions than removing data, such as fixing spelling and syntax errors, standardizing data sets, and correcting mistakes such as empty fields, missing codes, and identifying duplicate data points. Data cleaning is considered a foundational element of the data science basics, as it plays an important role in the analytical process and uncovering reliable answers. Most importantly, the goal of data cleaning is to create data sets that are standardized and uniform to allow business intelligence and data analytics tools to easily access and find the right data for each query.

Regardless of the type of analysis or data visualizations you need, data cleaning is a vital step to ensure that the answers you generate are accurate. When collecting data from several streams and with manual input from users, information can carry mistakes, be incorrectly inputted, or have gaps. Data cleaning helps ensure that information always matches the correct fields while making it easier for business intelligence tools to interact with data sets to find information more efficiently. One of the most common data cleaning examples is its application in data warehouses.

A successful data warehouse stores a variety of data from disparate sources and optimizes it for analysis before any modeling is done. To do so, warehouse applications must parse through millions of incoming data points to make sure they're accurate before they can be slotted into the right database, table, or other structure. Organizations that collect data directly from consumers filling in surveys, questionnaires, and forms also use data cleaning extensively. In their cases, this includes checking that data was entered in the correct field, that it doesn't feature invalid characters, and that there are no gaps in the information provided

## Data Analysis

Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.

A simple example of Data analysis is whenever we take any decision in our day-to-day life is by thinking about what happened last time or what will happen by choosing that particular decision. This is nothing but analyzing our past or future and making decisions based on it. For that, we gather memories of our past or dreams of our future. So that is nothing but data analysis. Now same thing analyst does for business purposes, is called Data Analysis.

### Types of Data Analysis: Techniques and Methods

There are several types of Data Analysis techniques that exist based on business and technology. However, the major Data Analysis methods are:

- Text Analysis
- Statistical Analysis
- Diagnostic Analysis
- Predictive Analysis
- Prescriptive Analysis

## Text Analysis

Text Analysis is also referred to as Data Mining. It is one of the methods of data analysis to discover a pattern in large data sets using databases or data mining tools. It used to transform raw data into business information. Business Intelligence tools are present in the market which is used to take strategic business decisions. Overall, it offers a way to extract and examine data and deriving patterns and finally interpretation of the data.

## Statistical Analysis

Statistical Analysis shows "What happen?" by using past data in the form of dashboards. Statistical Analysis includes collection, Analysis, interpretation, presentation, and modeling of data. It analyses a set of data or a sample of data. There are two categories of this type of Analysis - Descriptive Analysis and Inferential Analysis.

- Descriptive Analysis: analyses complete data or a sample of summarized numerical data. It shows mean and deviation for continuous data whereas percentage and frequency for categorical data.
- Inferential Analysis: analyses sample from complete data. In this type of Analysis, you can find different conclusions from the same data by selecting different samples.

## Diagnostic Analysis

Diagnostic Analysis shows "Why did it happen?" by finding the cause from the insight found in Statistical Analysis. This Analysis is useful to identify behavior patterns of data. If a new problem arrives in your business process, then you can look into this Analysis to find similar patterns of that problem. And it may have chances to use similar prescriptions for the new problems.

## Predictive Analysis

Predictive Analysis shows "what is likely to happen" by using previous data. The simplest data analysis example is like if last year I bought two dresses based on my savings and if this year my salary is increasing double then I can buy four dresses. But of course, it's not easy like this because you have to think about other circumstances like chances of prices of clothes is increased this year or maybe instead of dresses you want to buy a new bike, or you need to buy a house!

So here, this Analysis makes predictions about future outcomes based on current or past data. Forecasting is just an estimate. Its accuracy is based on how much detailed information you have and how much you dig in it.

### Prescriptive Analysis

Prescriptive Analysis combines the insight from all previous Analysis to determine which action to take in a current problem or decision. Most data-driven companies are utilizing Prescriptive Analysis because predictive and descriptive Analysis are not enough to improve data performance. Based on current situations and problems, they analyze the data and make decisions.

## Exploratory Data Analysis ~ EDA

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

1. maximize insight into a data set;
2. uncover underlying structure;
3. extract important variables;
4. detect outliers and anomalies;
5. test underlying assumptions;
6. develop parsimonious models; and
7. determine optimal factor settings.

The EDA approach is precisely that--an approach--not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.

EDA is not identical to statistical graphics although the two terms are used almost interchangeably. Statistical graphics is a collection of techniques--all graphically based and all focusing on one data characterization aspect. EDA encompasses a larger venue; EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model. EDA is not a mere collection of techniques; EDA is a philosophy as to how we dissect a data set; what we look for; how we look; and how we interpret. It is true that EDA heavily uses the collection of techniques that we call "statistical graphics", but it is not identical to statistical graphics per se.

### History of EDA

EDA holds its roots from the seminal work in EDA that is Exploratory Data Analysis, Tukey, (1977). Over the years it has benefitted from other noteworthy publications such as Data Analysis and Regression, Mosteller and Tukey (1977), Interactive Data Analysis, Hoaglin (1977), The ABC's of EDA, Velleman and Hoaglin (1981) and has gained a large following as "the" way to analyze a data set.

## Techniques of EDA

Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out.

The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

1. Plotting the raw data (such as data traces, histograms, bi-histograms, probability plots, lag plots, block plots, and Youden plots.
2. Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
3. Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

## Exploratory Data Analysis vs Classical Data Analysis (EDA vs CDA)

EDA is a data analysis approach. Besides EDA other data analysis approaches also exist and now question arises how does EDA differ from these other approaches. Three popular data analysis approaches are:

1. Classical
2. Exploratory (EDA)
3. Bayesian

These three approaches are similar in that they all start with a general science/engineering problem and all yield science/engineering conclusions. The difference is the sequence and focus of the intermediate steps.

- For classical analysis, the sequence is
  Problem → Data → Model → Analysis → Conclusions
- For EDA, the sequence is
  Problem → Data → Analysis → Model → Conclusions
- For Bayesian, the sequence is
  Problem → Data → Model → Prior Distribution → Analysis → Conclusions

Thus, for classical analysis, the data collection is followed by the imposition of a model (normality, linearity, etc.) and the analysis, estimation, and testing that follows are focused on the parameters of that model. For EDA, the data collection is not followed by a model imposition; rather it is followed immediately by analysis with a

goal of inferring what model would be appropriate. Finally, for a Bayesian analysis, the analyst attempts to incorporate scientific/engineering knowledge/expertise into the analysis by imposing a data-independent distribution on the parameters of the selected model; the analysis thus consists of formally combining both the prior distribution on the parameters and the collected data to jointly make inferences and/or test assumptions about the model parameters.

In the real world, data analysts freely mix elements of all of the above three approaches (and other approaches). The above distinctions were made to emphasize the major differences among the three approaches.

Focusing on EDA versus classical, these two approaches differ as follows:

1. Models
2. Focus
3. Techniques
4. Rigor
5. Data Treatment
6. Assumptions

## Data Interpretation

Data interpretation is the process of reviewing data through some predefined processes which will help assign some meaning to the data and arrive at a relevant conclusion. It involves taking the result of data analysis, making inferences on the relations studied, and using them to conclude.

Therefore, before one can talk about interpreting data, they need to be analyzed first. From the previous two sections we know that data analysis is the process of ordering, categorizing, manipulating, and summarizing data to obtain answers to research questions. It is usually the first step taken towards data interpretation. It is evident that the interpretation of data is very important, and as such needs to be done properly. Therefore, researchers have identified some data interpretation methods to aid this process.

### Methods involved in Data Interpretation

Data interpretation methods are how analysts help people make sense of numerical data that has been collected, analyzed and presented. Data, when collected in raw form, may be difficult for the layman to understand, which is why analysts need to break down the information gathered so that others can make sense of it. For example, when founders are pitching to potential investors, they must interpret data (e.g., market size, growth rate, etc.) for better understanding.

There are 2 main methods in which this can be done, namely; quantitative methods and qualitative methods.

## Qualitative Data Interpretation Method

The qualitative data interpretation method is used to analyze qualitative data, which is also known as categorical data. This method uses texts, rather than numbers or patterns to describe data. Qualitative data is usually gathered using a wide variety of person-to-person techniques, which may be difficult to analyze compared to the quantitative research method.

Unlike the quantitative data which can be analyzed directly after it has been collected and sorted, qualitative data needs to first be coded into numbers before it can be analyzed. This is because texts are usually cumbersome, and will take more time and result in a lot of errors if analyzed in its original state. Coding done by the analyst should also be documented so that it can be reused by others and also analyzed. There are 2 main types of qualitative data, namely; nominal and ordinal data. These 2 data types are both interpreted using the same method, but ordinal data interpretation is quite easier than that of nominal data.

In most cases, ordinal data is usually labelled with numbers during the process of data collection, and coding may not be required. This is different from nominal data that still needs to be coded for proper interpretation.

## Quantitative Data Interpretation Method

The quantitative data interpretation method is used to analyze quantitative data, which is also known as numerical data. This data type contains numbers and is therefore analyzed with the use of numbers and not texts. Quantitative data are of 2 main types, namely; discrete and continuous data. Continuous data is further divided into interval data and ratio data, with all the data types being numeric.

Due to its natural existence as a number, analysts do not need to employ the coding technique on quantitative data before it is analyzed. The process of analyzing quantitative data involves statistical modelling techniques such as standard deviation, mean and median. Some of the statistical methods used in analyzing quantitative data are highlighted below:

1. Mean: The mean is a numerical average for a set of data and is calculated by dividing the sum of the values by the number of values in a dataset. It is used to get an estimate of a large population from the dataset obtained from a sample of the population.
2. Standard deviation: This technique is used to measure how well the responses align with or deviates from the mean. It describes the degree of consistency within the responses; together with the mean, it provides insight into data sets.

3. Frequency distribution: This technique is used to assess the demography of the respondents or the number of times a particular response appears in research. It is extremely keen on determining the degree of intersection between data points.

Some other interpretation processes of quantitative data not used as popularly as the previous three and uses quite hold a small niche includes:

- Regression analysis
- Cohort analysis
- Predictive and prescriptive analysis

## Important points while collecting data for accurate data interpretation

- Identify the Required Data Type

  Researchers need to identify the type of data required for particular research. Is it nominal, ordinal, interval, or ratio data? The key to collecting the required data to conduct research is to properly understand the research question. If the researcher can understand the research question, then he can identify the kind of data that is required to carry out the research.

- Avoid Biases

  There are different kinds of biases a researcher might encounter when collecting data for analysis. Although biases sometimes come from the researcher, most of the biases encountered during the data collection process is caused by the respondent. There are 2 main biases, that can be caused by the President, namely; response bias and non-response bias. Researchers may not be able to eliminate these biases, but there are ways in which they can be avoided and reduced to a minimum.

  Response biases are biases that are caused by respondents intentionally giving wrong answers to responses, while non-response bias occurs when the respondents don't give answers to questions at all. Biases are capable of affecting the process of data interpretation.

- Use Close Ended Surveys

  Although open-ended surveys are capable of giving detailed information about the questions and allow respondents to fully express themselves, it is not the best kind of survey for data interpretation. It requires a lot of coding before the data can be analyzed. Close-ended surveys, on the other hand, restrict the respondents' answer to some predefined options, while simultaneously eliminating irrelevant data. This way, researchers can easily analyze and interpret data.

# Data Visualization

Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets. The term is often used interchangeably with others, including information graphics, information visualization and statistical graphics.

Data visualization is one of the steps of the data science process, which states that after data has been collected, processed and modeled, it must be visualized for conclusions to be made. Data visualization is also an element of the broader data presentation architecture (DPA) discipline, which aims to identify, locate, manipulate, format and deliver data in the most efficient way possible.

Data visualization is important for almost every career. It can be used by teachers to display student test results, by computer scientists exploring advancements in artificial intelligence (AI) or by executives looking to share information with stakeholders. It also plays an important role in big data projects. As businesses accumulated massive collections of data during the early years of the big data trend, they needed a way to quickly and easily get an overview of their data. Hence, visualization tools were a natural fit in these cases. Visualization is central to advanced analytics for similar reasons. When a data scientist is writing advanced predictive analytics or machine learning (ML) algorithms, it becomes important to visualize the outputs to monitor results and ensure that models are performing as intended. This is because visualizations of complex algorithms are generally easier to interpret than numerical outputs.

## Data visualization and big data

The increased popularity of big data and data analysis projects have made visualization more important than ever. Companies are increasingly using machine learning to gather massive amounts of data that can be difficult and slow to sort through, comprehend and explain. Visualization offers a means to speed this up and present information to business owners and stakeholders in ways they can understand. Big data visualization often goes beyond the typical techniques used in normal visualization, such as pie charts, histograms and corporate graphs. It instead uses more complex representations, such as heat maps and fever charts. Big data visualization requires powerful computer systems to collect raw data, process it and turn it into graphical representations that humans can use to quickly draw insights.

While big data visualization can be beneficial, it can pose several disadvantages to organizations. They are as follows:

- To get the most out of big data visualization tools, a visualization specialist must be hired. This specialist must be able to identify the best data sets and visualization styles to guarantee organizations are optimizing the use of their data.
- Big data visualization projects often require involvement from IT, as well as management, since the visualization of big data requires powerful computer hardware, efficient storage systems and even a move to the cloud.
- The insights provided by big data visualization will only be as accurate as the information being visualized. Therefore, it is essential to have people and processes in place to govern and control the quality of corporate data, metadata and data sources.

## Techniques involved in visualization

In the early days of visualization, the most common visualization technique was using a Microsoft Excel™ spreadsheet to transform the information into a table, bar graph or pie chart. While these visualization methods are still commonly used, more intricate techniques are now available, including the following:

- infographics
- bubble clouds
- bullet graphs
- heat maps
- fever charts
- time series charts

Some other popular techniques are as follows.

- Line charts. This is one of the most basic and common techniques used. Line charts display how variables can change over time.
- Area charts. This visualization method is a variation of a line chart; it displays multiple values in a time series -- or a sequence of data collected at consecutive, equally spaced points in time.
- Scatter plots. This technique displays the relationship between two variables. A scatter plot takes the form of an x- and y-axis with dots to represent data points.
- Treemaps. This method shows hierarchical data in a nested format. The size of the rectangles used for each category is proportional to its percentage of the whole. Treemaps are best used when multiple categories are present, and the goal is to compare different parts of a whole.
- Population pyramids. This technique uses a stacked bar graph to display the complex social narrative of a population. It is best used when trying to display the distribution of a population.

## Importance of Visualization

Data visualization provides a quick and effective way to communicate information in a universal manner using visual information. The practice can also help businesses identify which factors affect customer behavior; pinpoint areas that need to be improved or need more attention; make data more memorable for stakeholders; understand when and where to place specific products; and predict sales volumes.

Other benefits of data visualization include the following:

- the ability to absorb information quickly, improve insights and make faster decisions;
- an increased understanding of the next steps that must be taken to improve the organization;
- an improved ability to maintain the audience's interest with information they can understand;
- an easy distribution of information that increases the opportunity to share insights with everyone involved;
- eliminate the need for data scientists since data is more accessible and understandable; and
- an increased ability to act on findings quickly and, therefore, achieve success with greater speed and less mistakes

## Visualization Tools as of 2020

Data visualization tools provide data visualization designers with an easier way to create visual representations of large data sets. When dealing with data sets that include hundreds of thousands or millions of data points, automating the process of creating a visualization, at least in part, makes a designer's job significantly easier.

These data visualizations can then be used for a variety of purposes: dashboards, annual reports, sales and marketing materials, investor slide decks, and virtually anywhere else information needs to be interpreted immediately.

The best data visualization tools on the market have a few things in common. First is their ease of use. There are some incredibly complicated apps available for visualizing data. Some have excellent documentation and tutorials and are designed in ways that feel intuitive to the user. Others are lacking in those areas, eliminating them from any list of "best" tools, regardless of their other capabilities. The best tools can also handle huge sets of data. In fact, the very best can even handle multiple sets of data in a single visualization.

The best tools also can output an array of different chart, graph, and map types. Most of the tools below can output both images and interactive graphs. There are exceptions to the variety of output criteria, though. Some data visualization tools focus on a specific type of chart or map and do it very well. Those tools also have a

place among the "best" tools out there. Finally, there are cost considerations. While a higher price tag doesn't necessarily disqualify a tool, the higher price tag has to be justified in terms of better support, better features, and better overall value.

Notable honorariums go to the following tools

1. Power BI

   Power BI is a Business Intelligence and Data Visualization tool which helps you to convert data from various data sources into interactive dashboards and reports. It provides multiple software connectors and services.

   Features:
   - You can manage reports easily using SaaS solution.
   - Power BI gives you real-time updates on the dashboard.
   - It provides secure and reliable connection to your data sources in the cloud or on-premise.
   - Allows data exploration using natural language query
   - Offers feature for dashboard visualization regularly updated with the community.
   - It is easy to use hybrid configuration.

2. Tableau

   Tableau is a robust tool for visualizing data in a better way. You can connect any database to create understandable visuals. It enables you to share visualization with other people.
   Features:
   - Moderate speed with options to optimize and enhance the progress of an operation.
   - It provides extensive options to secure data without scripting.
   - This app comes with different versions, such as the Tableau server, cloud, and desktop.
   - It can be integrated with more than 250 applications.
   - Tableau helps you to resolve big data-related issues.

3. Qlik

   Qlik is a data visualization software which is used for converting raw data into knowledge. This software acts like a human brain which works on "association" and can go into any direction to search the answers.
   Features:
   - This tool supports various forms of data presentation.
   - It provides transparent reporting and scalability with data integration.
   - Automatically maintains data association.

- It offers the fast integration of data from various sources into a single application.
- Qlik provides data visualization in a meaningful and innovative way.
- It helps you to identify trends and information to make any decisions.

4. Adaptive Insights

   Adaptive Insights is a data visualization tool built to boost your business. It helps you to plan, budget, as well as forecast to make better decisions.

   Features:

   - You can easily collaborate with other people.
   - It automates data collection to ensure that you are working with fresh data.
   - This tool has a dashboard that enables you to create a report without effort.
   - Adaptive Insights enables you to plan a budget for a project.
   - You can adapt the plan in real time.

5. Dundas BI

   Dundas BI is an enterprise-ready Business Intelligence platform. You can deploy it as the central data portal for your company or integrate into any website.

   Features:

   - You can visually prepare and transform your business data into understandable reports.
   - It provides a vast range of layout options to choose from.
   - You can customize interactive charts, maps, and more.
   - Dundas helps you to incorporate data via an intuitive drag-and-drop facility.
   - Supports a wide range of statistical formulas.
   - Connect, interact, or analyze your data on any device.
   - It has an API that enables you to enhance design requirements the way your users desire.
   - Customizable data visualizations.

6. Domo

   Domo is a cloud platform that helps you to conduct analysis and create interactive visualization. It enables you to examine important data using graphs and pie charts. This app helps you to simplify administration data.

   Features:

   - You can get a real-time view of your data.

- Domo allows you to customize text, images, and color.
- You can see the visualization on any device.
- It enables you to set up customized alerts.
- This tool automatically monitors your data correlations, summaries, and more.
- You can integrate it with any software.
- Personalization of the dashboard is possible.

7. Cluvio

   Cluvio is a platform that enables you to run SQL queries for your database. It allows you to visualize the result in a better and understandable way.

   Features:

   - Cluvio helps you to translate your raw data into numerous professional charts and graphs.
   - A shared dashboard can be accessed by others without log in.
   - It allows you to share a dashboard with clients and colleagues.
   - Reports can be attached in an email.
   - You can filter multiple dashboards by any attributes of your data.
   - It automatically suggests a way to visualize data as a chart.
   - You can specify SQL alerts, a condition you like to be informed about.

8. Datawrapper

   Datawrapper is an open-source tool that enables you to create interactive charts. You can load CSV (Comma-separated Values data files into this app and embed maps onto your website.

   Features:

   - You can customize the app without writing any code.
   - Datawrapper supports Linux, Mac, and Windows operating systems.
   - It helps you to create maps, charts, and tables.
   - This tool enables you to connect your visualization to a Google Sheet.
   - You can use Datawrapper in any device.

9. Plotly

   Plotly is a tool that helps you to build analytic web apps. This app enables you to export HTML files and images in the report.

   Features:

   - It supports drag and drop functionalities.
   - You can easily search for a wide range of charts and templates.

- This program helps you to compose publication-quality figures without writing a single line of code.
- Plotly can be integrated with the existing workflow of your company.
- It helps you to manage privacy.

10. RAWgraphs

RAWGraphs is a data visualization app which makes the visual representation of any complicated data simple. It works with CSV and TSV (Tab Separated Values). This app helps you to embed charts directly into your web pages.

Features:

- You can copy-paste your data into RAWGraphs.
- Supports numerous visual layouts.
- It enables you to map your data dimensions visually to understand it.
- RAWgraphs provide immediate feedback on the visuals you have made.
- You can export your work in a PNG or SVG file.
- This framework allows you to open your output in any vector graphics editing tool.

11. Highcharts

Highcharts is a library that is written in JavaScript. It provides a simple way to add interactive charts to your web app or website.

Features:

- It is build using HTML 5.
- You can make a bar, column, pie chart, etc. for your website or mobile app
- It is compatible with mobile and tablets.
- Highcharts enables you to modify axes in a chart.

12. Visually

Visually is a platform for data visualization and infographics. It has a collection of numerous online contents. This app enables you to share data visualizations with others.

Features:

- It helps you make a website from infographics.
- You can turn your numbers into image-based visualization.
- It enables you to streamline the product design process.

13. Google Charts

Google Charts is an interactive cloud service that creates graphical charts from the information supplied by users. You can use it to make a simple line chart or complex hierarchical tree.

Features:

- It has a rich gallery of interactive charts.
- You can configure charts the way you like.
- It supports numerous web browsers.
- Charts and dashboards can be controlled from the dashboard.
- Connect to your data in real time.
- Google Charts is compatible with Android and iOS platforms.

14. Sisense

Sisense is a data visualization software that enables you to simplify complex data from multiple sources. This tool helps you to transform data into actionable applicable components or visualizations.

Features:

- It helps you to mash up data and create an analytics app.
- You can embed analytics anywhere with a customizable feature.
- You can deploy your work on the cloud using Windows or Linux.
- Sisense enables you to specify access rights to users.
- It provides user role-based security.
- You can recover your data anytime and safeguard against errors.
- Unify unrelated data into one centralized place
- This tool provides a drag-and-drop user interface.
- You can export data to Excel, CSV, PDF, and other formats

15. FusionCharts

FusionCharts is a JavaScript library for data visualization. It offers more than 2000 maps and 100+ charts. This library gives a range of customizable options to build charts from raw data.

Features:

- It enables you to integrate with JavaScript framework or server-side programming language.
- Fusioncharts helps you to build charts with real-world data.
- Supports all browsers, including IE 6,7, and 8.
- It provides ready-made dashboards to start your work.
- This library allows you to export the result in bulk.

16. TeamMate Analytics

TeamMate Analytics is a suite of 150+ Computer Aided Audit Tools. These apps allow you to perform data analysis for organizations or clients.

Features:

- It includes more than 200 audit tests.
- TeamMate enables you to convert reports, text files, and PDF to excel ready analysis.
- It helps you to represent your content in graphical form.
- This tool has an advanced visualizer that allows you to explore data geographically.
- You can get a detailed report of statistics.

17. Chartblocks

Chartblocks is an app that helps you to build charts. This cloud-based tool enables you to embed charts to any website. You can use it to customize any charts and synch with any data source.

Features:

- It provides dozens of editable charts.
- You can share charts on any social media website, including Facebook and Twitter.
- This tool enables you to import data from spreadsheets, databases, etc.
- You can create charts in any device having any screen size.

18. Ember Charts

Ember Charts is a charting library built-in JavaScript. It helps you to create a bar, pie, and many other editable charts.

Features:

- You can add legends, labels, tooltips, and mouseover effects.
- It provides automatic resizing of charts.
- This tool adjusts margin and padding in the visual diagrams.

19. Polymaps

Polymaps is a JavaScript library for creating interactive and dynamic maps with ease. It uses SVG (Scalable Vector Graphics) to display data.

Features:

- It has the capability to supports numerous types of visual presentations for raw data.
- This tool uses CSS to design your data.
- It provides a tile format to publish content.

20. Leaflet

Leaflet is an open-source data visualization tool that works efficiently across major mobile platforms and Desktop PCs. It can also be executed with the help of API.

Features:

- You can restyle it using CSS3.
- It enables you to zoom a specific area of the image using a double click.
- You can easily drag the marker to the specific location of the map.
- This program supports Firefox, Chrome, Opera, Safari, and more.
- It provides markups and popups.

21. Sigmajs

    Sigmajs is an online app that is made for creating a graph. This app helps you to customize your drawing. You can also publish the final result on any website.

    Features:

    - It supports touch and mouse.
    - You can write your own JavaScript functions.
    - This tool automatically deals with JSON (JavaScript Object Notation) to load and parse the file.
    - You can make changes and refresh the graph anytime you like.

22. Looker

    Looker is a data visualization platform that enables you to explore, analyse, and share analytics with ease. You can use this program to convert your data into useful diagrams.

    Features:

    - It provides a dashboard to analyze your data deeply.
    - You can filter individuals or groups dynamically.
    - Looker helps you to combine numerous types of charts.
    - You can visualize data with subtotal in tables.
    - This app provides a modern API to integrate workflow.

## A little intro the problem at hand

We are tasked with generating a Customer 360° profile for business to understand the kind of customers they are dealing with and the nature of their requirements thereby target their business to optimize returns.

The customer base we are dealing with is completely from the country of Mexico, North America and the business in question are all restaurants but rest assured all are from Mexico disregarding their specific place of establishment.

The restaurants (los restaurantes) as are obtained from web search and via Wayback Machines (https://archive.org/web/) and google.com search listing we get

that the places and the people are from a file called geoplaces2.csv and the other data is most probably originating from the UCI machine learning repository(https://archive.ics.uci.edu/ml/datasets/Restaurant+%26+consumer+data ).

The main task now stands is to generate information of the customer base by a Customer 360° view with the following steps at hand.

1. Clean Data and Process if possible (EDA tasks)
2. Create Graphical plots
3. Create Dashboards
4. Generate a meaningful story (telling any rise or fall in business)
5. Accumulate everything in a composite manner to give a 360-degree view of the customer database

# TABLEAU

Tableau Software is an American interactive data visualization software company focused on business intelligence. It was founded in Mountain View, California, and is currently headquartered in Seattle, Washington.

Tableau was founded in 2003 as a result of a computer science project at Stanford that aimed to improve the flow of analysis and make data more accessible to people through visualization. Co-founders Chris Stolte, Pat Hanrahan, and Christian Chabot developed and patented Tableau's foundational technology, VizQL—which visually expresses data by translating drag-and-drop actions into data queries through an intuitive interface.

Since our foundation, we've continuously invested in research and development at an unrivalled pace, developing solutions to help anyone working with data to get to answers faster and uncover unanticipated insights.

This includes making machine learning, statistics, natural language, and smart data prep more useful to augment human creativity in analysis. And we not only offer a complete, integrated analytics platform, but also proven enablement resources to help customers deploy and scale a data-driven culture that drives resilience and value through powerful outcomes.

Tableau was acquired by Salesforce in 2019, and our mission remains the same: to help people see and understand their data. Today, organizations everywhere—from non-profits to global enterprises, and across all industries and departments—are empowering their people with Tableau to drive change with data

Products:

- Tableau Desktop
- Tableau Server
- Tableau Online
- Tableau Prep Builder (Released in 2018)
- Tableau Vizable (Consumer data visualization mobile app released in 2015)
- Tableau Public (free to use)
- Tableau Reader (free to use)
- Tableau Mobile
- Tableau CRM

**NB: this project as told in during briefing – is supposed to be done by preferably Tableau Public or Tableau Desktop.

## Customer 360° view

The idea behind Customer 360 is to build a complete and accurate picture of every customer by aggregating all of each customer's structured and unstructured data from across your organization. With this unified knowledge, you can create wonderful customer experiences, personalize interactions with your customers, and build greater customer insights. The problem is that customer data exists in multiple databases across touchpoints (e.g., in-store, over the web, via social media, and telephone), geographies, and product lines. Duplicates and discrepancies are unavoidable when these systems are not synchronized.

Customer 360 acts as the hub that links and synchronizes the information about your customers. It becomes the source of reference for finding the most up-to-date information. Many call this the "single source of truth" about the customer. The data can then be de-duplicated, aggregated, analyzed, and displayed on demand. Customer 360 is not about creating a new database with all of your customer data. You do not replicate all of your data. Instead, you keep selected information in the Customer 360 database for fast access and then link the records to the source records in the source databases.

Customer 360 is also not about building a Single Customer View. Although, multiple Single Customer Views may be created from Customer 360 data. The biggest Customer 360 challenge is to correctly identify all of the records belonging to a particular customer because of inconsistencies across databases and poor data quality.

## How is Customer 360 different from Single Customer View?

While a Single Customer View (SCV) is sometimes called a 360-Degree Customer View, it is not the same as Customer 360. Unfortunately, the terms are very similar and confusion can result.

The distinction is that Single Customer View is about the presentation of customer information. It may be a single screen or report containing information about a specific customer that is aggregated from multiple sources. For example, a customer's ID number may be used to retrieve information stored separately on the sales database, finance database, and customer support database. The data itself may or may not be synchronized. SCV may just gather the customer information in a meaningful way to those who need it at the time they need it.

Different departments or individuals usually need a different view of the same customer. For example, a sales clerk may be given a view with a selected set of information and a customer service representative might be shown a different, but overlapping, set of information.

The key is that you can create a Single Customer View from selected customer data by pulling information from different databases without implementing Customer 360. This does have a risk that the data from different databases will not be synchronized and the data presented on the same screen could be inconsistent.

# Problem Statement

The Problem statement as was provided:

```
Project 2. A story with charts and dashboard to prepare 360° view of customer in tableau.

Use Tableau as a base to visualize from a customer database of an establishment based in Mexico and spanning across the country with data entries from different states of the country.
```

# General Idea of the Data Received

At first look we got four CSV (comma-separated variables) files namely

1. place details.csv
2. rating_final.csv
3. user order + payment1.csv
4. userprofile.csv

and as expectedly are expecting to contain the data the respective names are suggesting. As previously stated, the data is supposedly from UCI Machine Learning repository for Restaurants and Consumers used specifically for Restaurant suggestion (link:*https://archive.ics.uci.edu/ml/datasets/Restaurant+%26+consumer+data*)

Now for individual files since names are different it is implied that all files are considered in its originality. Details of each file specifically for this project is as follows.

1. Restaurant information (detailing of the place)
    - Filename: place details.csv;
    - Records: 131; Fields: 18;
    - Location: https://github.com/WolfDev8675/RepoSJX7/blob/Assign3_2/Data/place%20details.csv
2. Rating for the restaurant (multiple parameter rating by different users)
    - Filename: rating_final.csv;
    - Records: 1162; Fields: 5;
    - Location: https://github.com/WolfDev8675/RepoSJX7/blob/Assign3_2/Data/rating_final.csv
3. Orders placed and the Payments done
    - Filename: user order + payment1.csv;
    - Records: 1162; Fields: 6;
    - Location: https://github.com/WolfDev8675/RepoSJX7/blob/Assign3_2/Data/user%20order%20%2B%20payment1.csv
4. Profile of users in the loop or scope of data
    - Filename: userprofile.csv;
    - Records: 139; Fields: 19;
    - Location: https://github.com/WolfDev8675/RepoSJX7/blob/Assign3_2/Data/userprofile.csv

Primary filters and checks out files rating_final.csv and userprofile.csv to not contain any blanks for data, thereby removing the need for any operations on them.

## Problems in the Data

While working with the we found a lot of inconsistencies that needed to corrected before being used with Tableau itself also lies the fact that Tableau poorly handles empty or null datapoints in datasets.

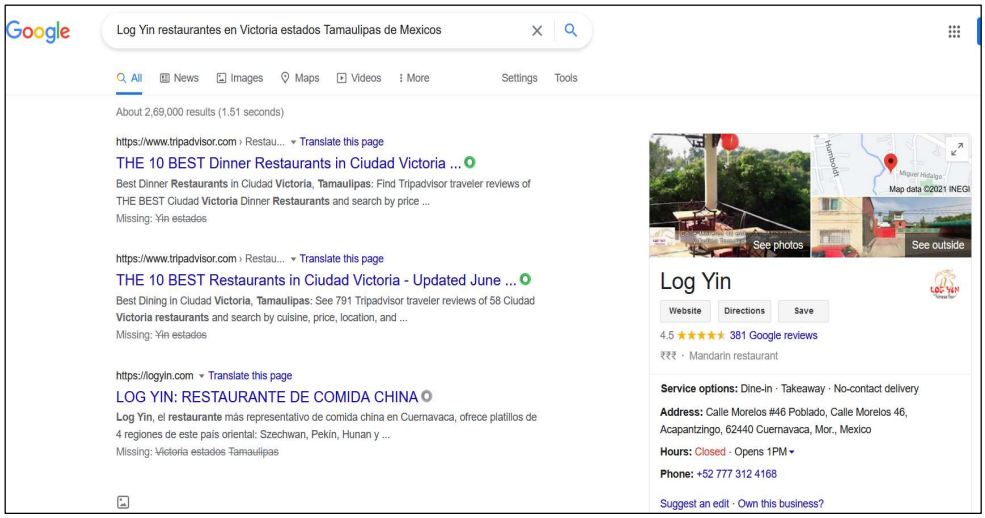
1. The data file dealing with Orders placed and payments had missing data values in the 'Billing amount' field 368 missing points out of 1161 records.

This problem was handled using the RANDBETWEEN function of Excel™ where the limits for this function is obtained from the minimum and maximum values from the rest of the data where the records aren't missing. This decision is taken owning to the fact the order quantity is from between 3 and 10 so even if we have variations in the rate itself, we also need to consider that there are various restaurants and the rate for the same food may be different from one establishment to other.

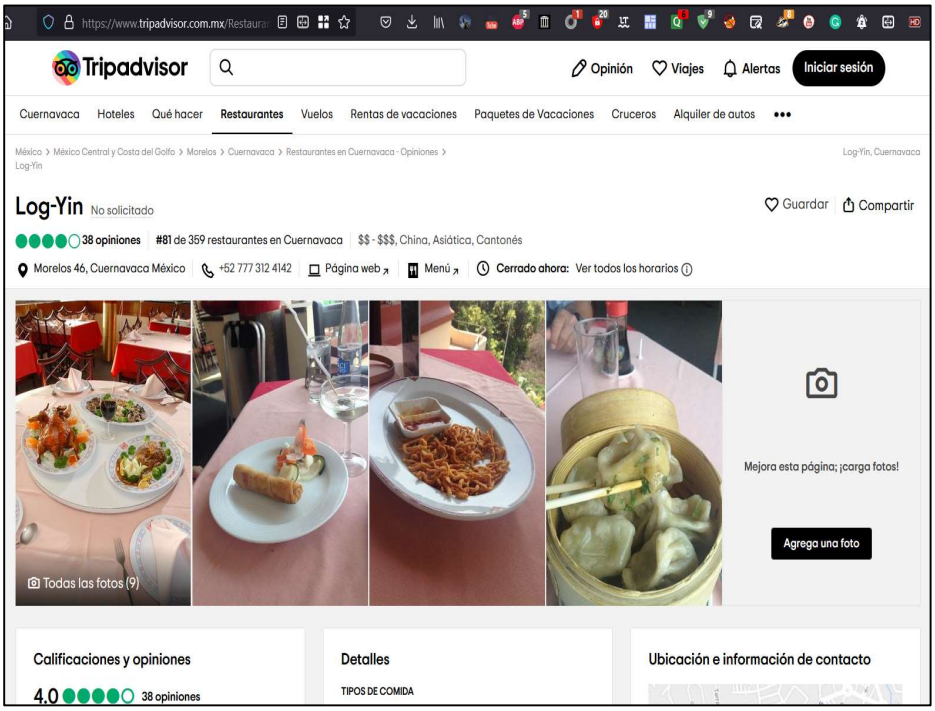
2. The data file dealing with places is more troubled than all other files.
    a. State names field had one instance of a record having nothing but a question mark ('?')
    b. There was another state field data of value 'mexico' where the country Mexico has no such state. There is a city though which is also the capital region of the country that is Mexico City.
    c. There are instances of 'Cd Victoria' (1 occurrence), 'Cd. Victoria' (3 occurrences) and 'Victoria' [lower caps] (11 occurrences) in the city field all belonging to the state 'Tamaulipas' where the name 'Tamaulipas' varied between lower and upper caps which interfered with xml base of tableau making all places different and the Mapbox extension of Tableau hit a dead-end trying to recognize these places in-spite of the places being described.
    d. Also, the Latitude and Longitude described in the data interfered with the placement.
    e. When primary mapped with tableu the places with the city Victoria was in fact placed far off the coast of Mexico. The Mapbox API recognizes the place to be Victoria Island in Lagos, Nigeria, but places the mark on the island of São Tomé et Príncipe, both of which lies in the western coastline of the African continent but separated by around 374 nautical miles.
    f. Beyond this lies specific random Restaurants which were searched directly via Mapbox online API and as well as on Google-Maps and Google-Earth that reveals places different from that on the map.

Inconsistencies in Google.com search:

Search keywords: "*Log Yin restaurantes en Victoria estados Tamaulipas de Mexicos*" reveals that there are no restaurants called Log Yin in Victoria, Tamaulipas, Mexico, this very Chinese restaurant lies in a different state of Mexico altogether let alone the city itself. Besides the location and address problems, all other attributes match, since the restaurant has a open space dining area, and all amenities as is described in the data record string.



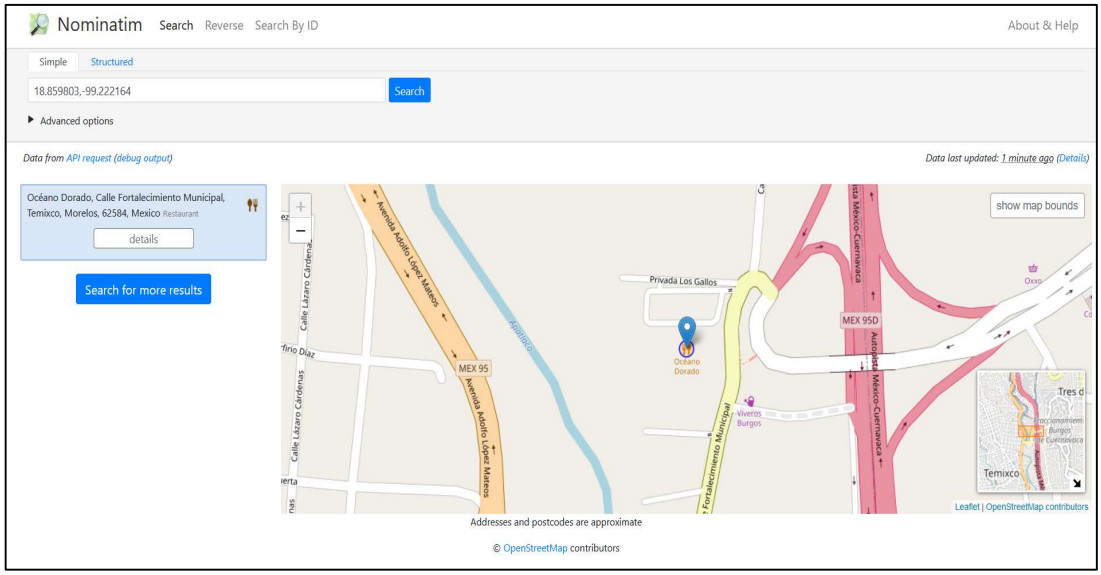

The same restaurant lies in Cuernavaca, Morelos




Although the Lat-Long pair is correct

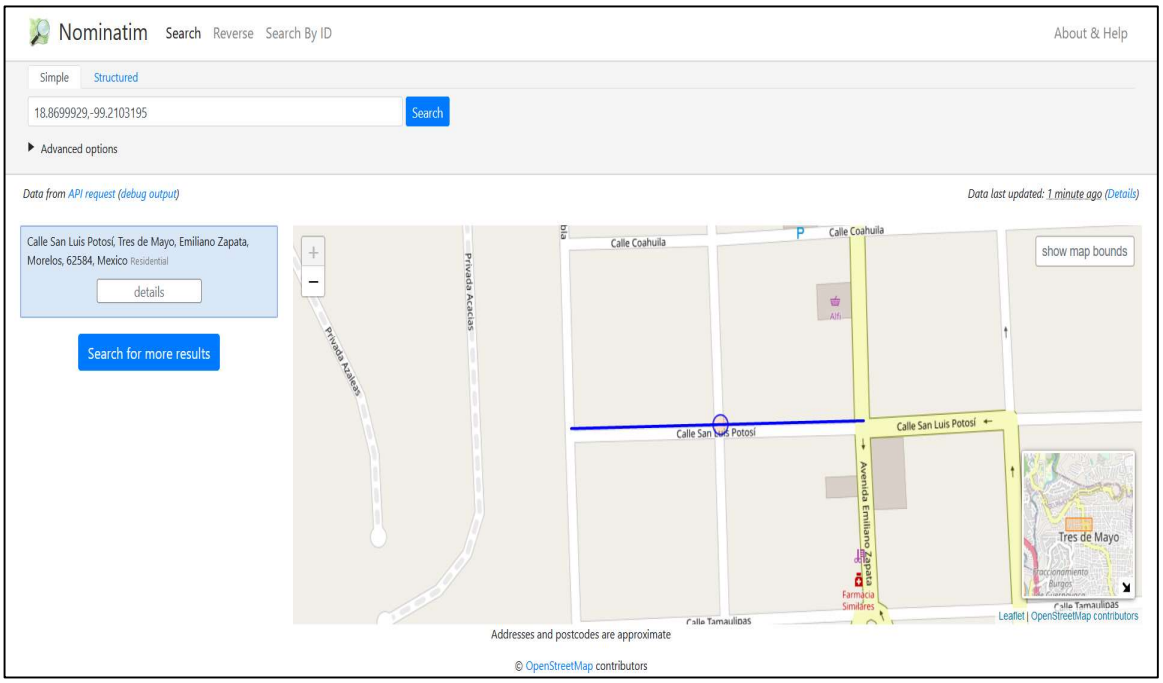
Similar search results are for places like Tortas Locas Hipocampo where the data file places it at San Luis Potosí but searches reveal restaurant of the same name exists in Puebla, east of Mexico City.

Restaurant "El Oceano Dorado" is said to be Cuernavaca, Morelos but in fact it lies in Temixco.



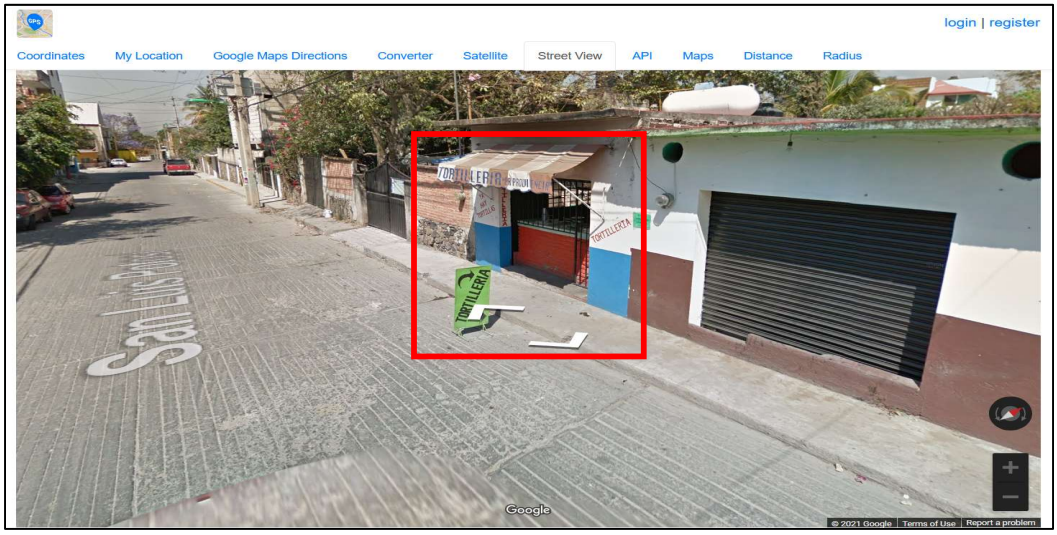

Another unique glitch noticed is of a place called El cotorreo with the record string as "`132773,18.8699929,-99.2103195,0101000020957F0000B1F935F9FA775AC1D06DFD1017514A41,El cotorreo,Emiliano Zapata Col. Tres de Mayo,Cuernavaca,Morelos,Mexico,Wine-Beer,permitted,informal,no_accessibility,low,familiar,f,open,live performance`" – GoogleSearch request: "*El Cotorro tres de Mayo en restaurante de Mexicos*". The latitude-longitude pair available is 18.8699929,-99.2103195 gives us a place where there is no place called El Cotorreo in Mexico but a restaurant called El Cotorro which is a Mexican Restaurant in Albuquerque, New Mexico, USA, not in Mexico basically changing the country altogether. But a place exists in the latitude-longitude pair.

Mapbox API gives us



Google Street-view gives us two variety of results one which match the four–way crossroad connector as given by Mapbox and the other just a few ten paces from the spot along the Calle San Luis Potosi (viz., San Luis Potosi Street) where there is a Tortileria (place for selling fresh Tortilla or a Tortilla bakery) that looks to be a small family-owned business which as of 2021 is permanently closed [*last noted open in 2019*]. (Street-View link: https://www.gps-coordinates.net/street-view/@18.869832,-99.210080,h270,p0,z3 )

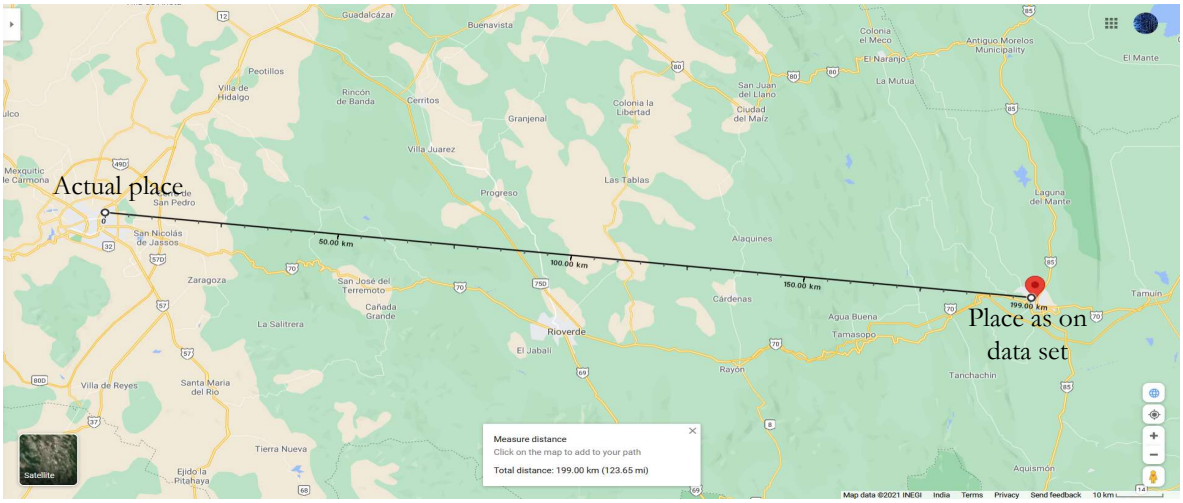
The crossing between Calle San Luis Potosi and Emiliano Zapata Avenue in Tres de Mayo, Morelos as given in the panoramic image below. In fact, all of the establishments in an around this crossroad connector are all home-improvement places varying from home décor, ceramic and tiles, paints and color, etc., except for a lone servicing place for automobiles. Just as an example the region has within a circle of 50m has at least 3 ceramic (cerámica) shops.

The Tortileria itself, a place which barely looks to be an open restaurant or a place which can host live performance as is described in the data.

Another significant glitch noted is for the place called Rincon Huasteco, data record information: "`132830,22.1508494,-100.9397522,0101000020957F0000F2E21C48374958C1C6826887BA984B41,Rincon Huasteco,Zaragoza entre Francisco Zarco y Lopez Velarde,Cd. Victoria,Tamaulipas,Mexico,No_Alcohol_Served,Anywhere,informal,completely,low,familiar,f,closed,live performance`" – depicts a place to exist in Ciudad Victoria, Tamaulipas. Google search (Search request: "*Rincon Huasteco en restaurante de Mexicos*") of the place reveals a location located in San Luis Potosi not in Tamaulipas.

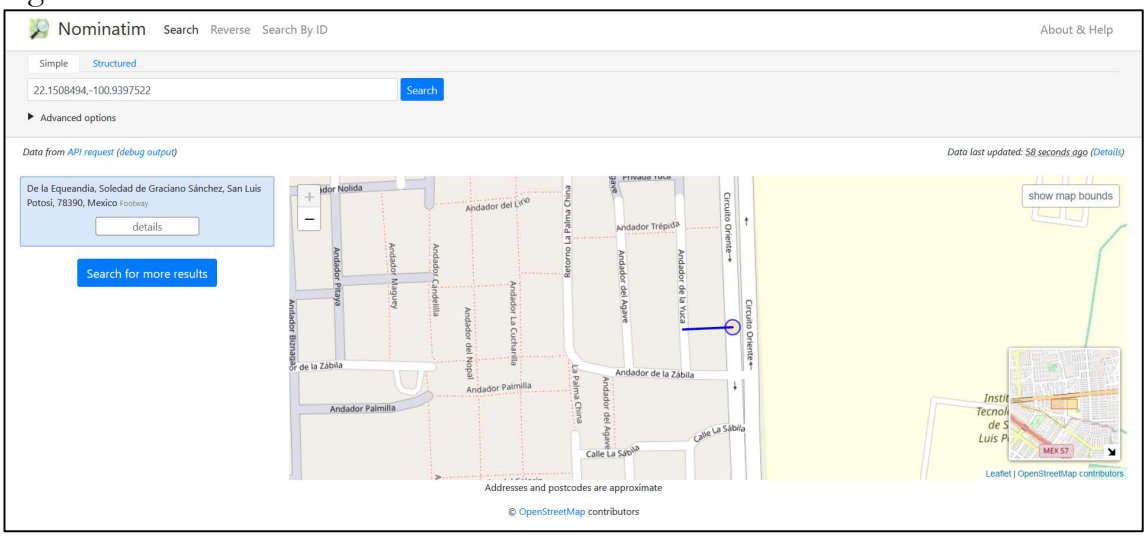

Contrary to this location the latitude-longitude pair (22.1508494, -100.9397522) gives us a location in Soledad de Graciano Sánchez, San Luis Potosi a distance of 199km by a crow's flight.



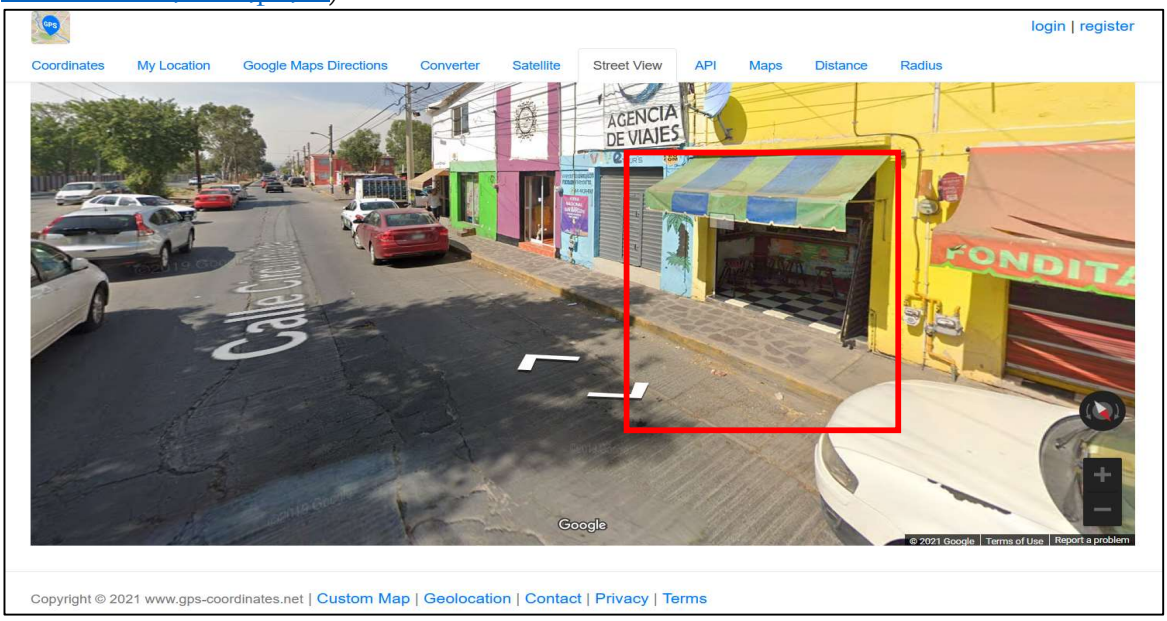

Mapbox results for the latitude-longitude pair (22.1508494, -100.9397522) depicts a place opposite the compound of Technological Institute of San Luis Potosi

(Instituto Tecnológico de San Luis Potosí) as given below, and supposedly the far west parking lot of the institute is in the East end and the place we noted.

If traversed along this street (Calle Circuito Ote) north – south movement there are scattered small scale restaurants distributed at various places around this region.




The Google Street-view of the same location (Street View link: https://www.gps-coordinates.net/street-view/@22.1508494,-100.9397522,h270,p0,z3)




A location which looks to be some business establishment but not the one we are looking for (viz., *Rincon Huestaco*) in fact most current map data gives us the same place to be a convenience store of the name 'La Surianita'. Even the neighbouring establishment's (Fondita La Abuelita[ESP] aka Granny Fondita[EN-GB]) information is also missing as of 2021.

### The necessary evil

As noticed from the discrepancies in the place names and the location depicted in their corresponding latitude-longitude pairs. Since GPS location mainly detects the location address rather than any establishment existing at that place, also lies a fact that any establishment could be closed down, the owners may move their business elsewhere or the shop may be sold out to another person who brings a different genre of business to the place. Considering this problem with the name irrecoverable cause either the name will fail or the latitude-longitude pair fails where in such a case one such information needs to be scrapped and preference over a place-name to Lat-Long pair is not briefed to us ever while the project was dispatched or said in later discussions we take the Lat-Long pair to correct the zonal address (Post-code, City, State, Country, Pin-code, etc.) than checking whether the place has physical existence in that very spot. Also, if considered 131 establishments could not be checked on individual basis and fixed with the places where they lie.

### The Solution

The very probable solution taken to combat this problem to some extent is to code using GPS APIs to correct the Zonal location giving preference to the Latitude-Longitude given in the data string and correct the place address but in the process leaving out the crosschecking the existence of the specific place in that specific location.

## The Correcting Code

For correcting the zonal location as we set out for, we used the OpenStreetMap contributing APIs that has MapBox support and Nominatim API and the GeoPy module for geocoders. The reason for using the Nominatim API is mainly due to cost related reasons. GeoPy has maps.google API extension as well as Mapbox API extension both of which requires a secure key that needs to be purchased on the basis of 'pay as you go', although Nominatim API has a premium version which uses a secure key access that has more accuracy than the free, non–premium version, but the free version is sufficient for the task we set for as well as the request limit is well within the number of places we need to check and correct. The code for editing place is as given below:

PlaceEdits.py

```
#!usr/bin/python

""" Code for editing place details using GPS locations
 for file 'place details.csv' """
# code file: PlaceEdits.py
# code author: BishalBiswas(https://github.com/WolfDev8675)
```

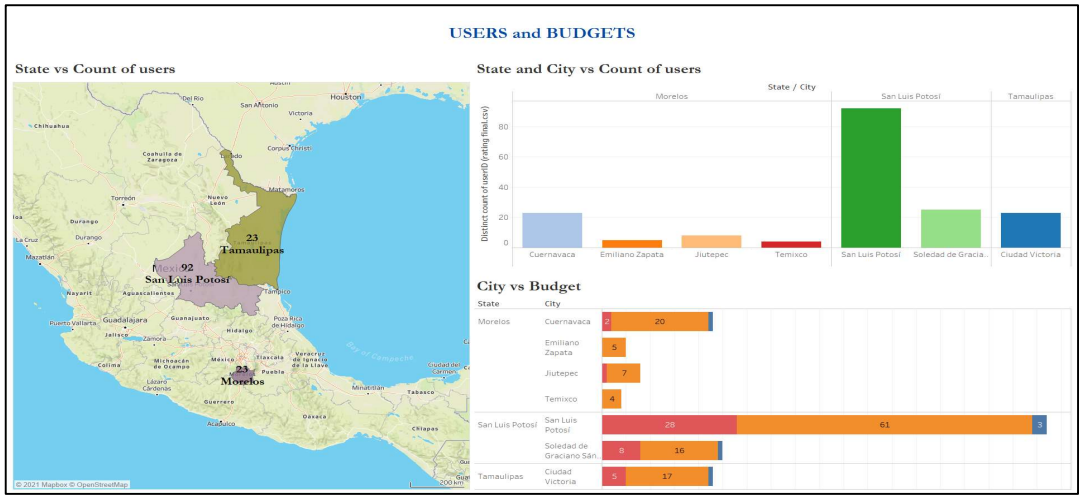
```
#imports
import pandas as PD
from geopy.geocoders import Nominatim as G_Loc


# DataFrame
placeDF=PD.read_csv('e:\Source\Repos\WolfDev8675\RepoSJX7\Data\place
details.csv',encoding='latin-1')

# operate
idx=placeDF.index.to_list()
GeoLoc=G_Loc(user_agent='Generic')
for one_id in idx:
  lat=str(placeDF.loc[one_id,'latitude'])
  lon=str(placeDF.loc[one_id,'longitude'])
  plc=GeoLoc.reverse(lat+","+lon).raw['address']

  try:
      placeDF.loc[one_id,'city']=plc['city']
  except:
      placeDF.loc[one_id,'city']=plc['county'] # since counties are found to be city
equivalent matching state outlines perfectly in this case
      # the classical case of city regions being shared between city municipality and
another smaller neighbouring municipality

#push to secondary source
placeDF.to_csv("e:\Source\Repos\WolfDev8675\RepoSJX7\EditedData\place_details.csv",
encoding='latin-1', index=False)


#end of code
```

## A look into the presentable 360-degree analysis

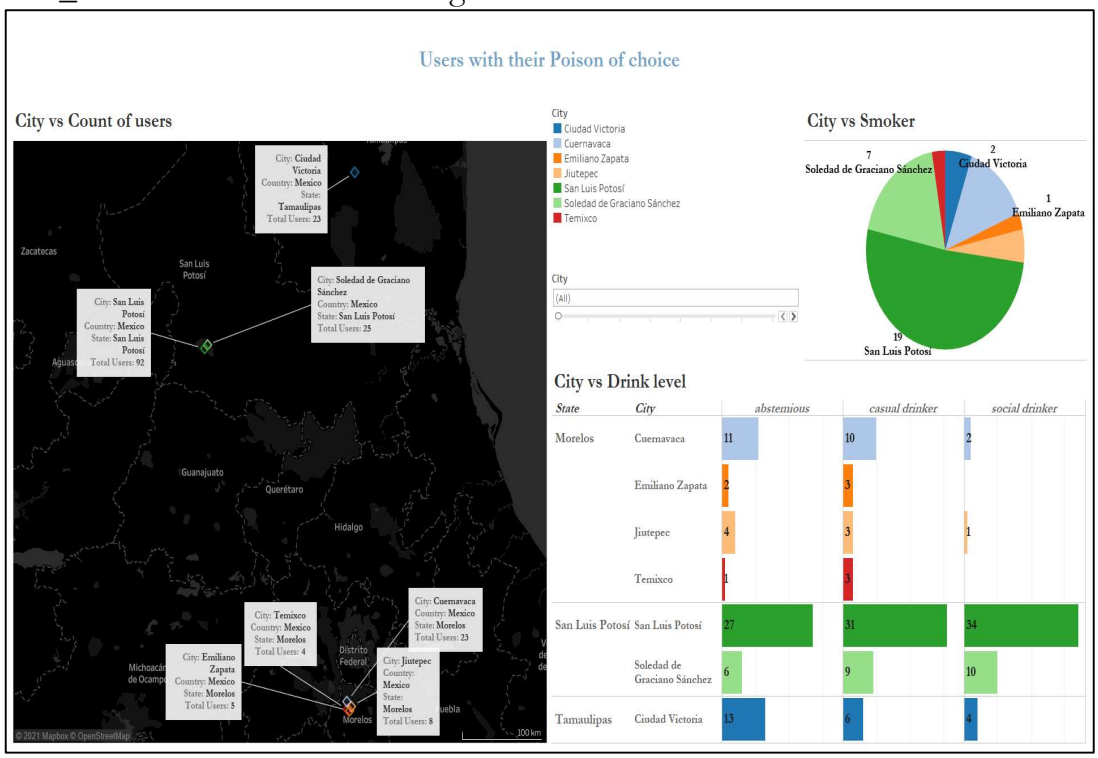
Considering the current stage of the data after the corrections we made into them to be final and correct. The following interactive dashboards were obtained, that has information about the business nature.

### Dashboards dealing with user-profile

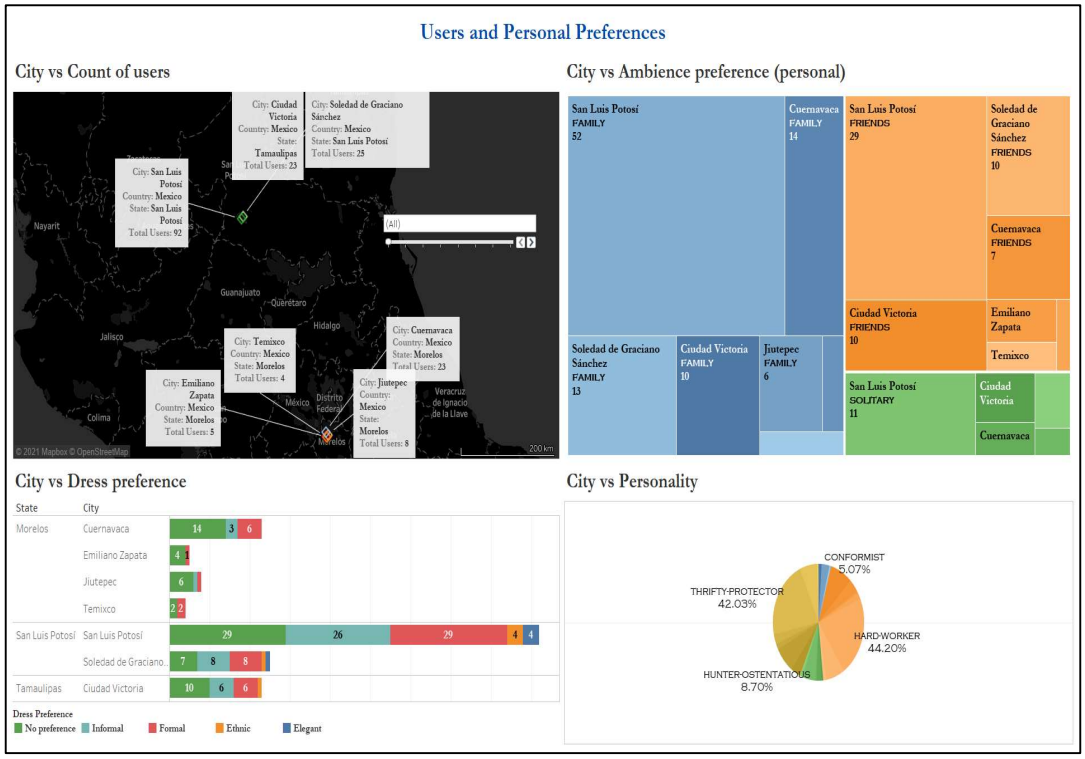
1. UPr_dsh1: Dealing with the user base and their budget limits.



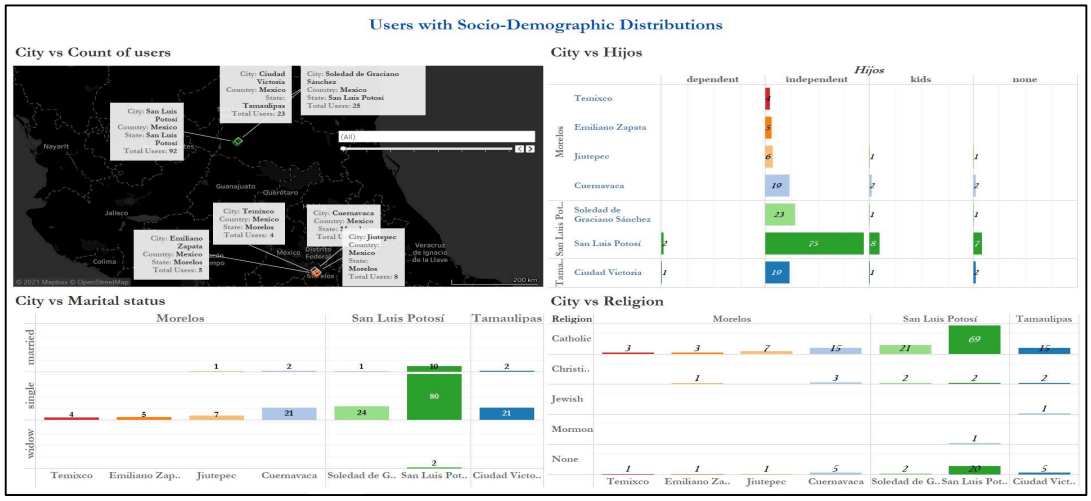
2. UPr_dsh2: Smokers and Drinking nature.




3. UPr_dsh3: Dealing with Personal preferences.

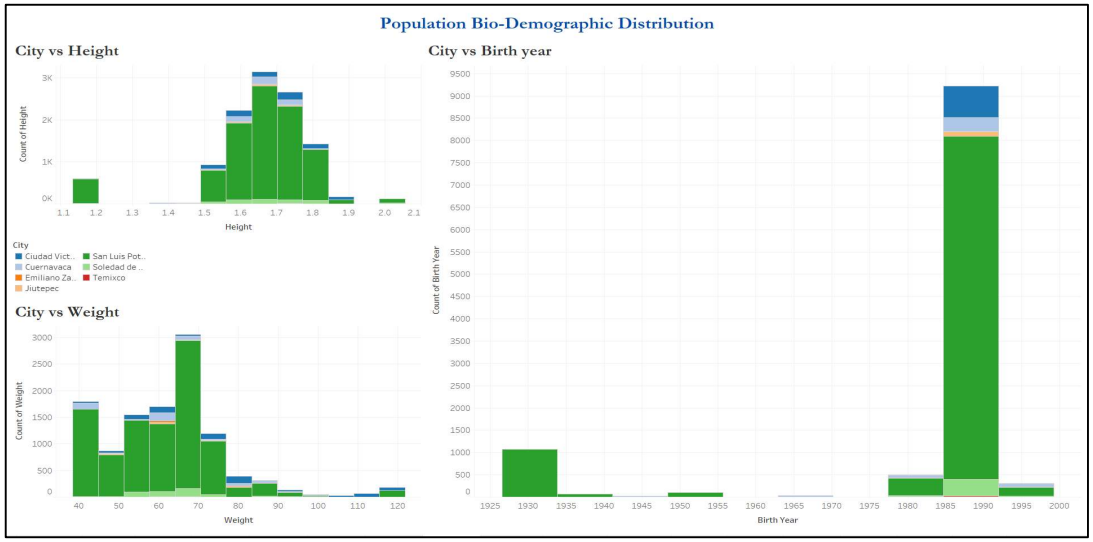


4. UPr_dsh4: Social Demographics.



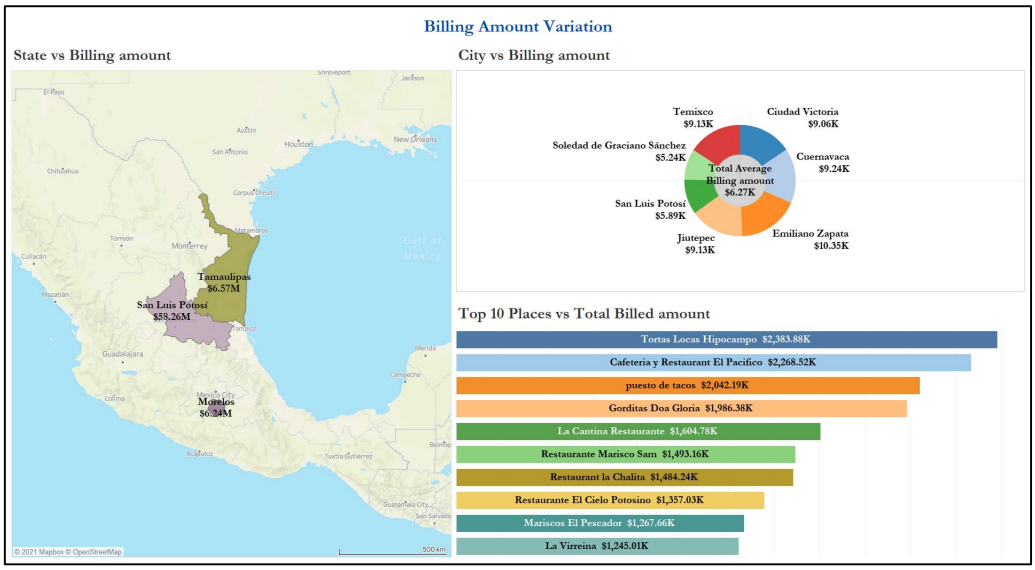

5. UPr_dsh5: Socio – Political study.


6. UPr_dsh6: Age & Health statistics.

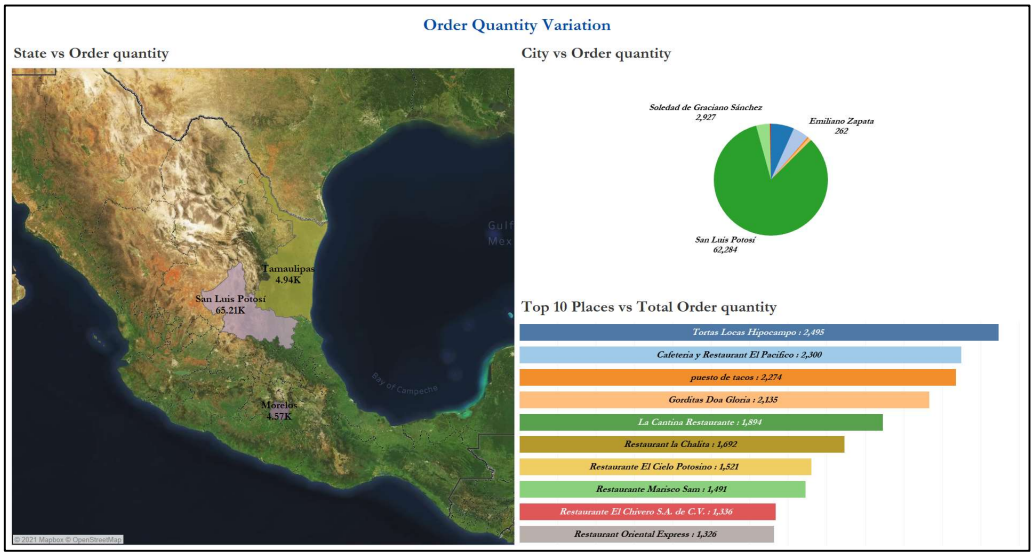


Dashboards dealing with user – order and payments

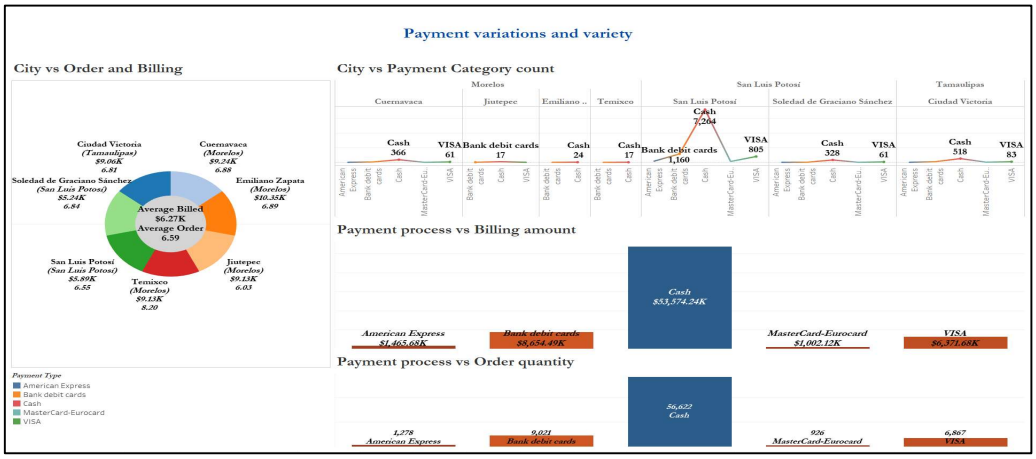
1. UoPy_dsh1: Variation of billing amount with place.




2. UoPy_dsh2: Variation of order quantity with place.

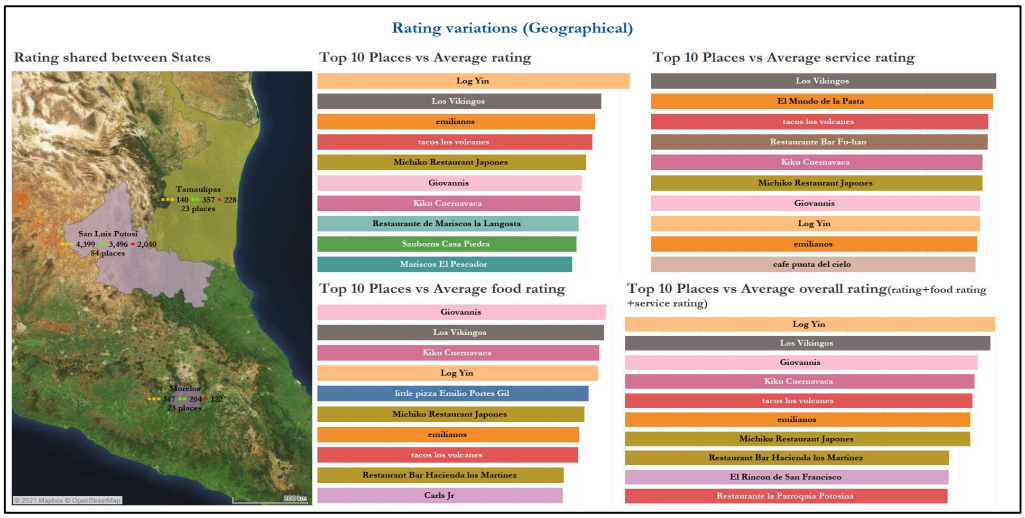



3. UoPy_dsh3: Payment process variation with place.

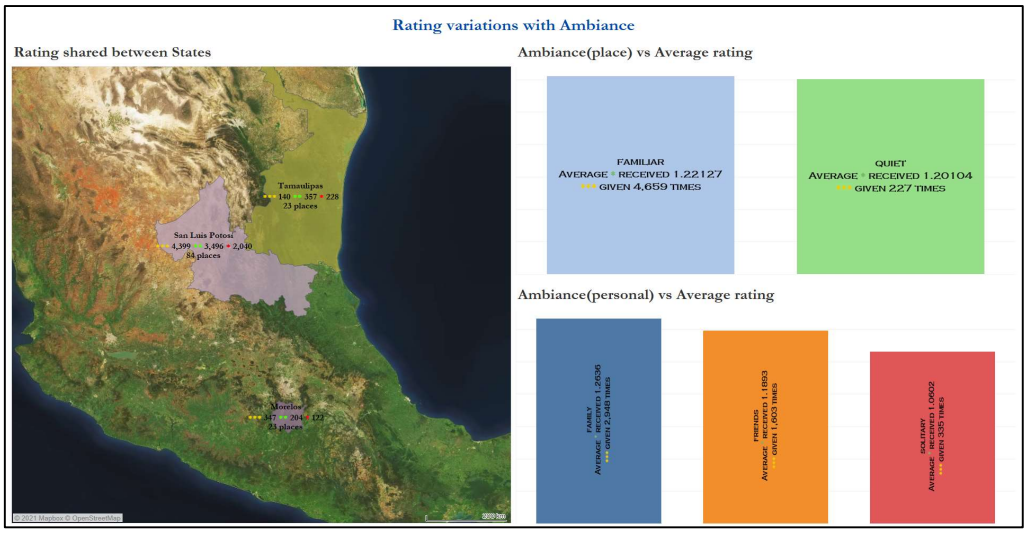


## Dashboards dealing with ratings

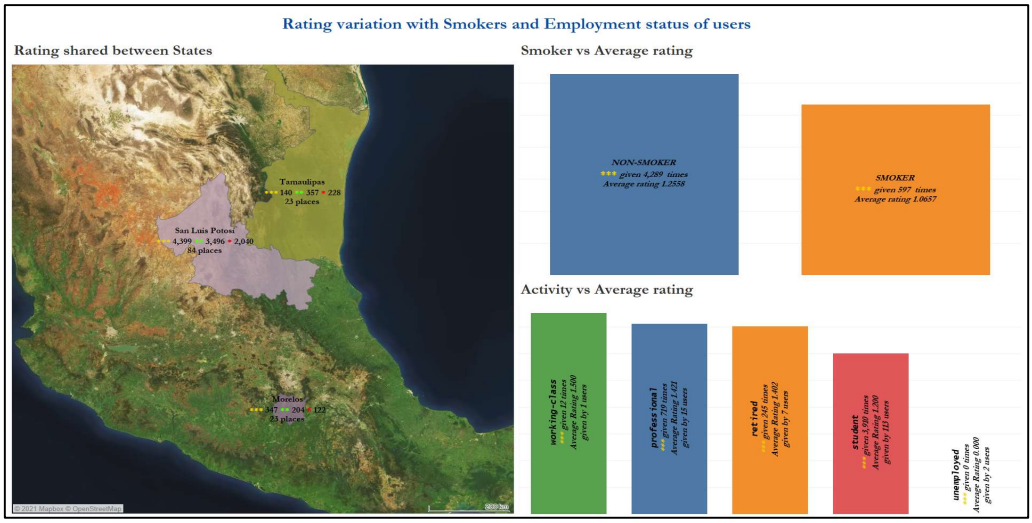
1. RtF_dsh1: Variation of rating with respect to place.



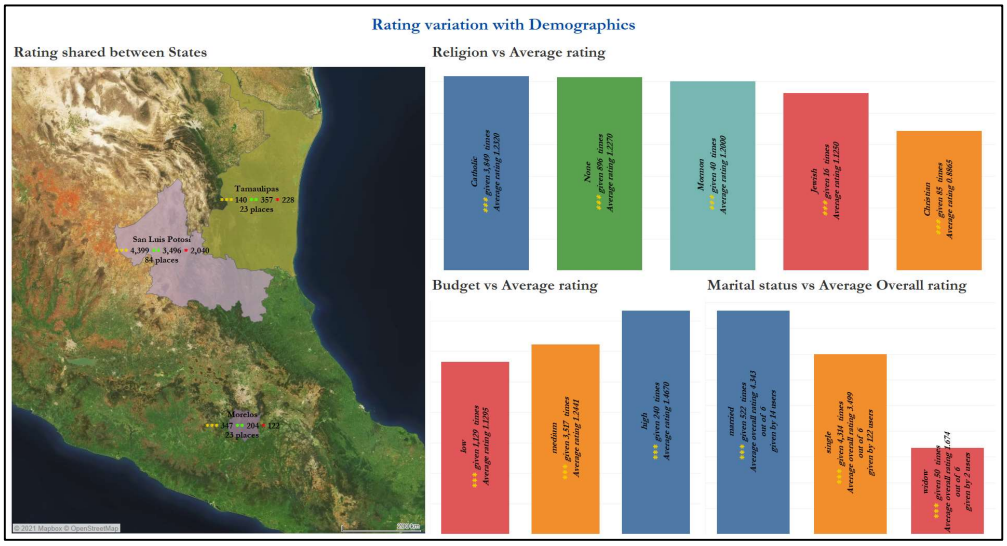

2. RtF_dsh2: Variation of rating with respect to ambiance.




3. RtF_dsh3: Variation of rating with respect to smokers and employment status of customers.

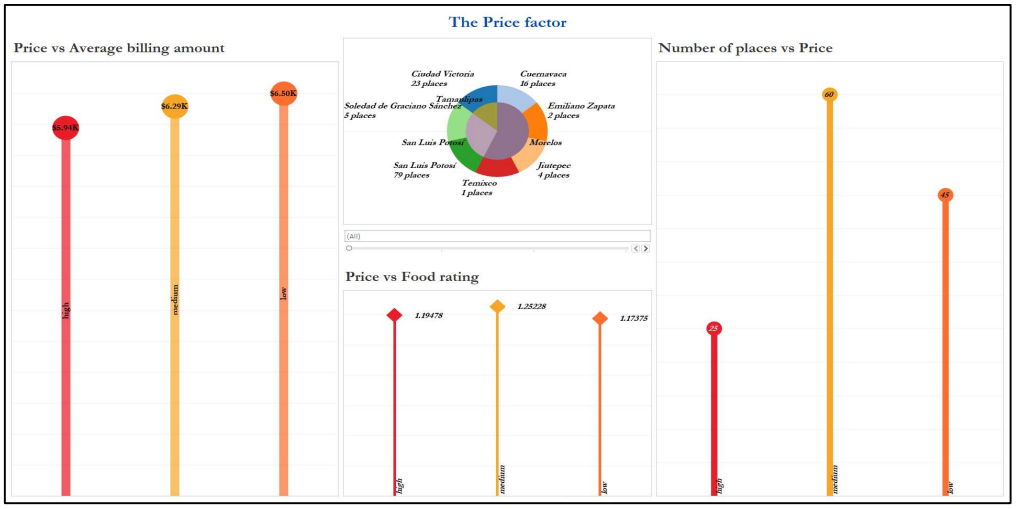


4. RtF_dsh4: Variation of rating with Demographics.

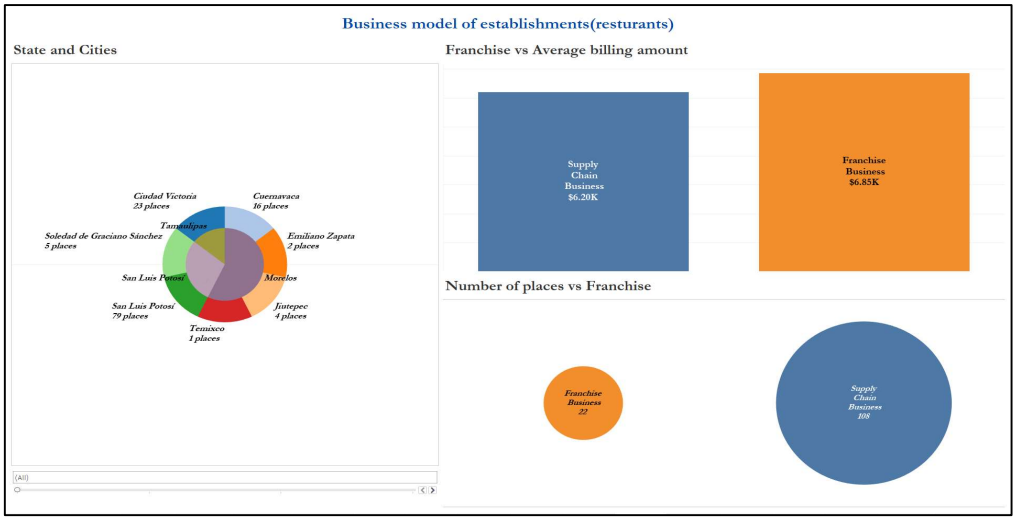



## Dashboards dealing with place – details

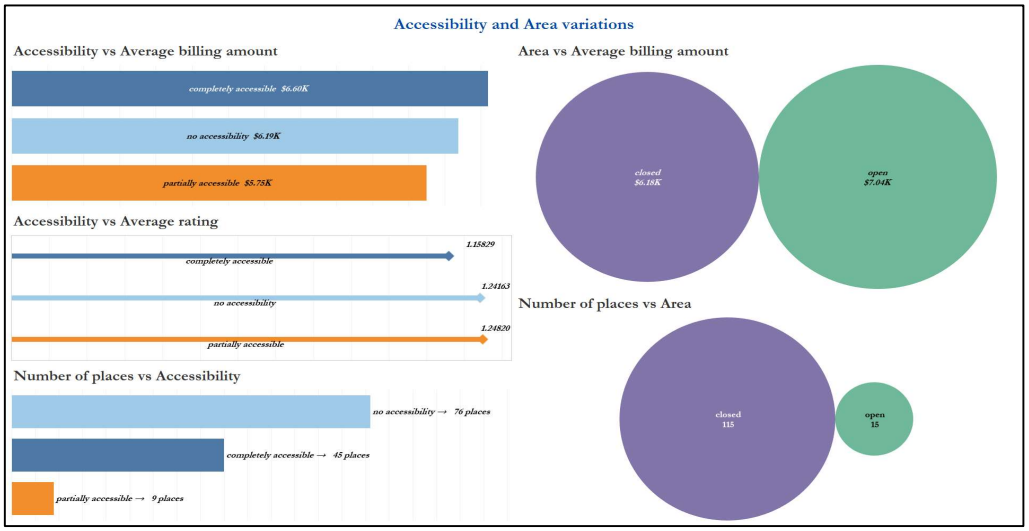
1. PlD_dsh1: Pricing details by geographical locations.



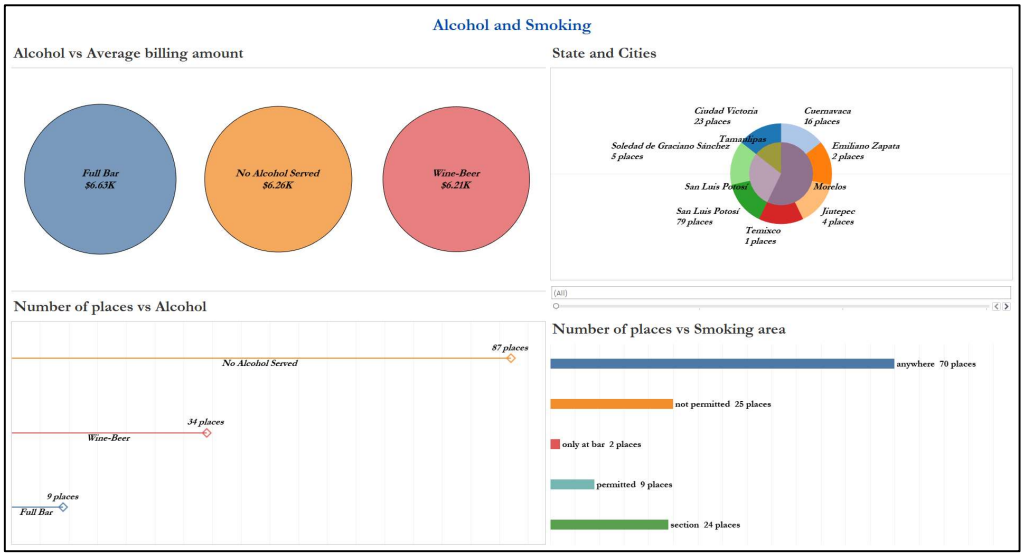

2. PlD_dsh2: Business models followed.



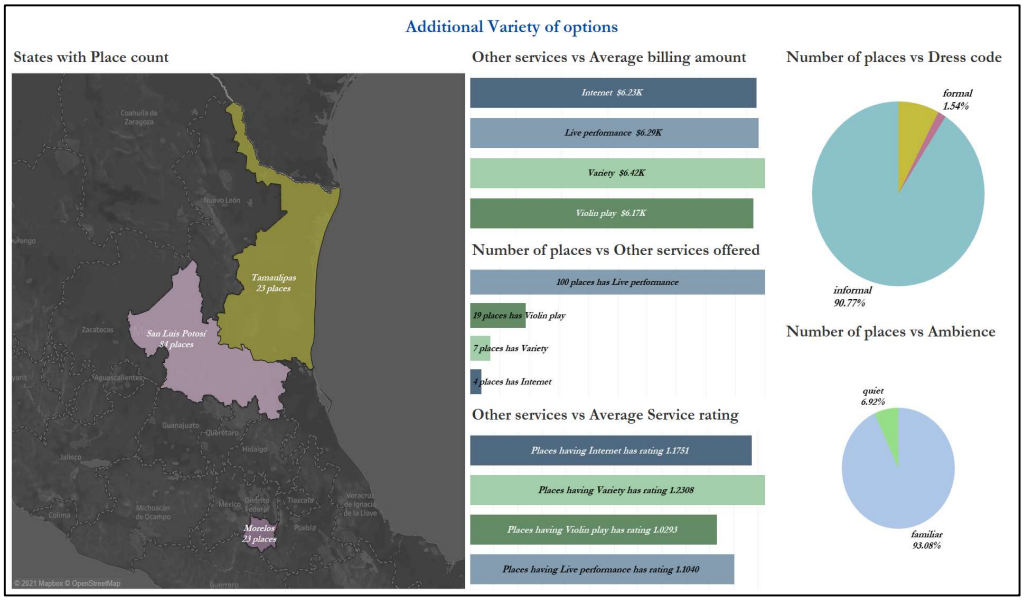
3. PlD_dsh3: Study of accessibility and area of establishments.




4. PlD_dsh4: Alcohol and Smoking restrictions.




5. PlD_dsh5: Miscellaneous options provided with respect to places.



## Chart lists

All the charts as is graphed in the project workbook of tableau is as given below. For convenience the charts are sectioned into groups with respect to their governing primary datasheet (although not directly relevant but used for proper search and backtracking worksheets during dashboard and story creation)

Charts under user – profile definitions

| Sl. No. | Chart name | Description | Used in | Chart type |
|---|---|---|---|---|
| 1 | UPr_Pac1 | State vs Count of users | UPr_dsh1 | Map |
| 2 | UPr_Pac2 | City vs Count of users | UPr_dsh2, UPr_dsh3, UPr_dsh4, UPr_dsh5 | Map |
| 3 | UPr_Pac3 | State and City vs Count of users | UPr_dsh1 | Bar |
| 4 | UPr_Pac4 | City vs Smoker | UPr_dsh2 | Pie |
| 5 | UPr_Pac5 | City vs Drink level | UPr_dsh2 | Bar |
| 6 | UPr_Pac6 | City vs Dress preference | UPr_dsh3 | Bar (multi) |
| 7 | UPr_Pac7 | City vs Ambience (personal) | UPr_dsh3 | Tree blocks |
| 8 | UPr_Pac8 | City vs Transport | UPr_dsh5 | Donut |
| 9 | UPr_Pac9 | City vs Marital Status | UPr_dsh4 | Bar |
| 10 | UPr_Pac10 | City vs Hijos | UPr_dsh4 | Bar |
| 11 | UPr_Pac11 | City vs Interest | UPr_dsh5 | Bar |
| 12 | UPr_Pac12 | City vs Personality | UPr_dsh3 | Pie |
| 13 | UPr_Pac13 | City vs Religion | UPr_dsh4 | Bar |
| 14 | UPr_Pac14 | City vs Activity | UPr_dsh5 | Donut |
| 15 | UPr_Pac15 | City vs Color | UPr_dsh5 | Tree blocks |
| 16 | UPr_Pac16 | City vs Budget | UPr_dsh1 | Bar |
| 17 | UPr_Pac17 | City vs Height | UPr_dsh6 | Histogram |
| 18 | UPr_Pac18 | City vs Weight | UPr_dsh6 | Histogram |
| 19 | UPr_Pac19 | City vs Birth year | UPr_dsh6 | Histogram |

Charts under orders and payments

| Sl. No. | Chart name | Description | Used in | Chart type |
|---|---|---|---|---|
| 1 | UoPy_Pac1 | State vs Billing amount | UoPy_dsh1 | Map |
| 2 | UoPy_Pac2 | State vs Order quantity | UoPy_dsh2 | Map |
| 3 | UoPy_Pac3 | City vs Billing amount | UoPy_dsh1 | Donut |
| 4 | UoPy_Pac4 | City vs Order quantity | UoPy_dsh2 | Pie |
| 5 | UoPy_Pac5 | City vs Payment Category count | UoPy_dsh3 | Line |

| | | | | |
|---|---|---|---|---|
| 6 | UoPy_Pac6 | Payment process vs Order quantity | UoPy_dsh3 | Bar |
| 7 | UoPy_Pac7 | Payment process vs Billing amount | UoPy_dsh3 | Bar |
| 8 | UoPy_Pac8 | Top 10 places vs Total Billed Amount | UoPy_dsh1 | Bar |
| 9 | UoPy_Pac9 | Top 10 places vs Total Order quantity | UoPy_dsh2 | Bar |
| 10 | UoPy_Pac10 | City vs Order and Billing | UoPy_dsh3 | Donut |

Charts under rating details

| Sl. No. | Chart name | Description | Used in | Chart type |
|---|---|---|---|---|
| 1 | RtF_Pac1 | Top 10 places vs Average rating | RtF_dsh1 | Bar |
| 2 | RtF_Pac2 | Top 10 places vs Average food rating | RtF_dsh1 | Bar |
| 3 | RtF_Pac3 | Top 10 places vs Average service rating | RtF_dsh1 | Bar |
| 4 | RtF_Pac4 | Top 10 places vs Average overall rating | RtF_dsh1 | Bar |
| 5 | RtF_Pac5 | Ambiance (place) vs Average rating | RtF_dsh2 | Bar |
| 6 | RtF_Pac6 | Religion vs Average rating | RtF_dsh4 | Bar |
| 7 | RtF_Pac7 | Smoker vs Average rating | RtF_dsh3 | Bar |
| 8 | RtF_Pac8 | Ambiance (personal) vs Average rating | RtF_dsh2 | Bar |
| 9 | RtF_Pac9 | Budget vs Average rating | RtF_dsh4 | Bar |
| 10 | RtF_Pac10 | Activity vs Average rating | RtF_dsh3 | Bar |
| 11 | RtF_Pac11 | Marital status vs Average overall rating | RtF_dsh4 | Bar |
| 12 | RtF_Pac12 | Rating shared between States | RtF_dsh1, RtF_dsh2, RtF_dsh3, RtF_dsh4 | Map |

Charts under place details

| Sl. No. | Chart name | Description | Used in | Chart type |
|---|---|---|---|---|
| 1 | PlD_Pac1 | Price vs Average billing amount | PlD_dsh1 | Lollipop |
| 2 | PlD_Pac2 | Franchise vs Average billing amount | PlD_dsh2 | Bar |

| | | | | |
|---|---|---|---|---|
| 3 | PlD_Pac3 | Other services vs Average billing amount | PlD_dsh5 | Bar |
| 4 | PlD_Pac4 | Accessibility vs Average billing amount | PlD_dsh3 | Bar |
| 5 | PlD_Pac5 | Alcohol vs Average billing amount | PlD_dsh4 | Sized Circles |
| 6 | PlD_Pac6 | Area vs Average billing amount | PlD_dsh3 | Sized Circles |
| 7 | PlD_Pac7 | Number of places vs Alcohol | PlD_dsh4 | Arrows |
| 8 | PlD_Pac8 | Number of places vs Smoking area | PlD_dsh4 | Bar |
| 9 | PlD_Pac9 | Number of places vs Dress code | PlD_dsh5 | Pie |
| 10 | PlD_Pac10 | Number of places vs Accessibility | PlD_dsh3 | Bar |
| 11 | PlD_Pac11 | Number of places vs Price | PlD_dsh1 | Lollipop |
| 12 | PlD_Pac12 | Number of places vs Ambience | PlD_dsh5 | Pie |
| 13 | PlD_Pac13 | Number of places vs Franchise | PlD_dsh2 | Sized Circles |
| 14 | PlD_Pac14 | Number of places vs Area | PlD_dsh3 | Sized Circles |
| 15 | PlD_Pac15 | Number of places vs Other services offered | PlD_dsh5 | Bar |
| 16 | PlD_Pac16 | Price vs Food rating | PlD_dsh1 | Lollipop |
| 17 | PlD_Pac17 | Accessibility vs Average rating | PlD_dsh3 | Bar |
| 18 | PlD_Pac18 | Other services vs Average Service rating | PlD_dsh5 | Bar |
| 19 | PlD_Pac19 | State and Cities;<br>States with place count | PlD_dsh1,<br>PlD_dsh2,<br>PlD_dsh4,<br>PlD_dsh5(Map) | Multi – level Pie;<br>Map |

## Critical understanding achieved from the project

Details and studies to extensive depths on the data of 180 customers gave us information on the customer base in quite a detail if not all since we don't have information all people. But to the extent collected and assimilated we have the following points to our disposal.

- Major portion of customers are middle class families and middle budget patrons of restaurants.

- Most of the restaurants also match up to this and are also medium cost establishments.
- High-cost restaurants have patrons rating food higher.
- San Luis Potosí has the largest of our customers we are dealing with.
- Smoking is usually not a restriction in most places but some few countable places have restrictions.
- Few places offer beer but there are also fewer Full Bar restaurants, but most places don't serve any alcoholic drinks.
- Dressing formal is liked by most people in San Luis Potosí.
- Majority of the customers are Students and also hard-working.
- Family and Friends are the people of choice when it comes to visiting restaurants.
- A Familiar environment is better chosen than other factors although liking varies and overall stays the same at average quality.
- Morelos is best for its food. Restaurants as Log Yin, Los Vikingos, Giovanni are must visit in Morelos.
- Payment methods tell that Temixco and Emiliano Zapata are smaller and lesser urbanized than the cities like Cuernavaca, Ciudad Victoria, San Luis Potosí.

# Bibliography

The following materials were referenced for bringing the project to fruition.


- https://www.lotame.com/what-are-the-methods-of-data-collection/
- https://www.guru99.com/what-is-data-analysis.html
- https://www.investopedia.com/terms/d/data-analytics.asp
- https://web.archive.org/web/20201205235717/https://www.ibm.com/in-en/analytics/hadoop/big-data-analytics
- https://www.talend.com/resources/what-is-data-preparation/
- https://www.zarantech.com/blog/importance-of-data-science/
- https://www.trifacta.com/data-cleansing/
- https://www.sisense.com/glossary/data-cleaning/
- https://www.analyticsvidhya.com/blog/2020/07/knnimputer-a-robust-way-to-impute-missing-values-using-scikit-learn/
- https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825
- https://www.import.io/post/business-data-analysis-what-how-why/
- https://www.itl.nist.gov/div898/handbook/eda/eda.htm
- https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

- https://www.formpl.us/blog/data-interpretation
- https://searchbusinessanalytics.techtarget.com/definition/data-visualization
- https://www.tableau.com/learn/articles/data-visualization
- https://www.guru99.com/best-data-visualization-tools.html
- https://archive.org/web/
- https://archive.ics.uci.edu/ml/datasets/Restaurant+%26+consumer+data
- https://en.wikipedia.org/wiki/Tableau_Software
- https://www.tableau.com/why-tableau/what-is-tableau
- https://globalz.com/customer-360-single-customer-view/
- https://www.sketchbubble.com/en/presentation-360-customer-profile.html


* * *

Complete collection of the project files is safely kept at

https://github.com/WolfDev8675/RepoSJX7/tree/Assign3_2

* * *