[Exploratory Data Analysis 3](#_Toc193832765)

[Project Part 1 3](#_Toc193832766)

[Project Description 3](#_Toc193832767)

[Project Title 3](#_Toc193832768)

[Objective: 3](#_Toc193832769)

[Dataset: 3](#_Toc193832770)

[Methodology: 4](#_Toc193832771)

[Tools and Technologies: 8](#_Toc193832772)

[Deliverables: 9](#_Toc193832773)

[Descriptive Analysis 9](#_Toc193832774)

[Week 6 Class Activity 9](#_Toc193832775)

[Project Description: 9](#_Toc193832776)

[Project Title: 9](#_Toc193832777)

[Objective: 10](#_Toc193832778)

[Dataset: 10](#_Toc193832779)

[Methodology: 11](#_Toc193832780)

[Tools and Technologies: 15](#_Toc193832781)

[Deliverables: 15](#_Toc193832782)

[Data Wrangling 16](#_Toc193832783)

[Project Part 1 16](#_Toc193832784)

[Project Description: Data Wrangling for Relations within Public-Trees 16](#_Toc193832785)

[Project Title: 16](#_Toc193832786)

[Objective: 16](#_Toc193832787)

[Background: 16](#_Toc193832788)

[Dataset: 16](#_Toc193832789)

[Methodology: 17](#_Toc193832790)

[Tools and Technologies: 20](#_Toc193832791)

[Deliverables: 20](#_Toc193832792)

[Data Quality Control 21](#_Toc193832793)

[Project Description: Data Quality Control Initiative for City of Vancouver 21](#_Toc193832794)

[Project Title: 21](#_Toc193832795)

[Objective: 21](#_Toc193832796)

[Background: 21](#_Toc193832797)

[Scope: 21](#_Toc193832798)

[Methodology: 22](#_Toc193832799)

[Cleaning: Week 3 Class Activity 27](#_Toc193832800)

[Profiling: Week 3&4 Activity 28](#_Toc193832801)

[Tools and Technologies: 30](#_Toc193832802)

[Deliverables: 30](#_Toc193832803)

[Timeline: 30](#_Toc193832804)

[Course Completion Badge 30](#_Toc193832805)

# Exploratory Data Analysis

## Project Part 1

Project Description: Exploratory Data Analysis (EDA) on Public-Trees Dataset for City of Vancouver

Project Title: Design and Implement DAP for City of Vancouver

Objective: The primary goal of this project is to perform an exploratory data analysis (EDA) on the Public-Trees dataset to discover whether the various data of the tree are related. By analyzing the columns such as Height\_Range\_ID, Height\_Range and Diameter, we are willing to know the relationship between these conditions.

Dataset: Downloaded from City of Vancouver Data Portal, Filtered “Acutissima” only for Species\_Name.

Figure 1

***Raw Dataset Download Filter***

A screenshot of a computer

AI-generated content may be incorrect.

*Note.* This screenshot is to show the filter of dataset on COV data portal. Source: City of Vancouver Data Portal.

* TREE\_ID: ID for trees
* CIVIC\_NUMBER: Civic number for trees
* STD\_STREET: Specific Street name
* GENUS\_NAME: Genus name for trees (Quercus only)
* SPECIES\_NAME: Species name for trees (Acutissima only)
* COMMON\_NAME: Common name for trees (Sawtooth Oak only)
* ON\_STREET\_BLOCK: Street block number
* ON\_STREET: Specific Street area name
* NEIGHBOURHOOD\_NAME: Specific neighbourhood name
* STREET\_SIDE\_NAME: Street side planted condition (Even/Odd)
* HEIGHT\_RANGE\_ID: ID for different Height Range (1, 2, 3, 4, 5, 6, 7)
* HEIGHT\_RANGE: The height range for trees (10-20, 20-30, 30-40, 40-50, 50-60, 60-70, 70-80)
* DIAMETER: Diameters for trees
* DATE\_PLANTED: The date of trees planted
* Geom: JSON format data storing geographic coordinates
* geo\_point\_2d: Easier to use storing geographic coordinates

## Methodology:

1. ***Data Collection and Preparation:***
   1. Download the PublicTrees dataset on City of Vancouver Data Portal as csv. Format
   2. Upload it to the correct folder in raw data bucket: I create a raw bucket, also create the folder for raw dataset.

Figure 2

***Upload data to Raw Bucket***

A screenshot of a computer

AI-generated content may be incorrect.

*Note*. This screenshot shows the raw dataset stored in the raw bucket. Source: AWS S3

* 1. Perform initial data profiling in AWS Glue Data Brew: Use AWS Glue Data Brew to run the data profiling.

Figure 3

***Data Profiling***

A screenshot of a computer

AI-generated content may be incorrect.

*Note*. This screenshot shows the profiling result of raw dataset in Data Brew. Source: AWS Glue Data Brew.

* 1. Do the cleaning after profiling to make sure data is ready to use: The dataset does not contain a lot of problems, I only changed the title of columns to make them look clearer. And I changed the format of data planted to yyyy-mm-dd, also to make sure it looks clearer.

Figure 4

***Data Cleaning***

A screenshot of a computer

AI-generated content may be incorrect.

*Note*. This screenshot shows the cleaning step of raw dataset in Data Brew. Source: AWS Glue Data Brew.

1. ***Descriptive Statistics: Project Part 2 Athena Business Question 1 and 2***
   * Business question 1: To see what’s the average diameter for first 20 rows.
   * Business question 2: To see if there’s a relationship between height range and diameter, by analyzing the average diameter grouped by different height range for first 20 rows.

Figure 5

***Athena Business Question 1***

A screenshot of a computer

AI-generated content may be incorrect.

*Note*. This screenshot shows the first business question of project part 2 in AWS Athena. Source: AWS Athena.

Figure 6

***Athena Business Question 2***

A screenshot of a computer

AI-generated content may be incorrect.

*Note*. This screenshot shows the second business question of project part 2 in AWS Athena. Source: AWS Athena.

1. Diameter Analysis:
   * Compare average diameter between different height range.
2. Insights and Findings:
   * From the results, we can see that the diameter is positively correlated with the height range of the tree, that is, the larger the height range of the tree, the larger the diameter will be.
3. Conclusion:
   * We can use the conclusion of this simple business question: the higher the height range, the larger the diameter to help predict future tree planting plans. People can use this conclusion to choose tree planting locations in a more planned way.

## Tools and Technologies:

* AWS S3
* AWS Glue Data Brew
* AWS Glue
* AWS Athena
* Excel

## Deliverables:

Exploratory data analysis is a very common category in our daily data analysis. It includes the preparation of the dataset in the early stage, the subsequent cleaning, the establishment of hypotheses and the use of tools to analyze and finally draw conclusions. We can apply the final conclusions to some scenarios. In general, this is a good exercise to help us prepare qualified datasets and use exploratory thinking to discover problems and try to find answers.

# Descriptive Analysis

## Week 6 Class Activity

Project Description:

These are Descriptive analysis of:

* Sales Performance Analysis
* Impact of Marketing Spend
* Effect of Temperature on Sales
* Website Traffic Insights
* Customer Purchase Behavior
* Outlier Detection

Project Title: Understanding These Business Questions based on Sales SSOT.

Objective: This exercise is to help us master how to convert business questions into data questions and implement them. In this process, we use the Sales\_SSOT dataset as an example. We have six categories and a total of 18 business questions for practice. The ultimate goal is to verify whether these assumptions are valid or obtain corresponding numerical conclusions.

Dataset: The dataset includes sales data, temperature data, market data, customer behavior data, website traffic data and Outlier flag data of SSOT from the same periods as January to March for 2023 and 2024, containing the following key features:

* Date: Date of data collection (2023/1/1-2023/3/31 & 2024/1/1-2024/3/31)
* Sales: The sales data on that day
* Temperature: Recorded temperature on that day
* Marketing\_Spend: The marketing spends on that day
* Customer\_Purchase\_Behavior: The customer behavior type (A, B, C)
* Website\_Traffic: The website traffic data on that day
* Outlier\_Flag: If there are outlier flags appear (0, 1)

Figure 7

***Sales-SSOT***

A screenshot of a computer

AI-generated content may be incorrect.

*Note*. This screenshot shows the table of dataset for Sales\_SSOT in Week 6 class activity. Source: UCW

## Methodology:

1. Data Collection and Cleaning:

Make sure the data looks clean and ready to use for analysis.

1. Understand the dataset and ask Business Questions

Understand what the different columns represent (mentioned in the dataset section above) and come up with some business questions.

1. Understand Business Questions and transform them to Data Questions

Deeply understand the business questions raised and convert them into data questions that we can use data to draw conclusions

Figure 8

***Business Question to Data Question***

A screenshot of a computer

AI-generated content may be incorrect.

*Note.* This is the screenshot of Business questions and data questions in Week 6 class activity. Source: UCW

1. Data Visualization:

Make the conclusions visual so that people who watch our project can more easily understand what we want to express.

Figure 9

***Visualizations of Business Questions***A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

*Note*. This screenshot shows the visualizations of different business questions in excel of Week 6 Class Activity.

1. Insights and Findings:

Draw some insights and conclusions and think about how to apply them to other/future scenarios, so that our analysis will have practical significance.

* 1. Example 1: Sales Performance Analysis: This visualization consists of total sales and dates. The general trend is upward, and the special thing is that the dates are arranged in the order of January 2023 and 2024 - February 2023 and 2024 - March 2023 and 2024.
  2. Example 2: Impact of Marketing Spend: This is a pie chart consisting of two dates and the corresponding sales on those days. The two dates are 2023/1/1 and 2023/3/15, representing the dates of the lowest and highest sales respectively.
  3. Example 3: Effect of Temperature on Sales: This is a bar chart, consisting of average sales and temperature. Average sales here represent the average of all sales at the corresponding temperature (such as 5, 6, 8, etc.).

1. Summarize the insights derived from the analysis, highlighting:
   1. Sales Performance Analysis: We can see from this visualization that as time goes by from January to March, total sales are increasing (and this happens in two years at the same time).
   2. Impact of Marketing Spend: We can conclude that sales are usually the lowest at the beginning of the year, and as time goes by until March, sales are likely to reach a peak.
   3. Effect of Temperature on Sales: We can see from the bar chart that there is a clear relationship between the two times, and the conclusion is that average sales will increase as the temperature increases.
2. Recommendations:
   1. Based on Sales Performance Analysis and Impact of Marketing Spend: This organization can consider using a strategy of gradually increasing output/sales from January to March each year to meet market demand. This can also be reflected in labor costs, because the market demand is low in January, and the company can consider reducing labor costs in January to reduce expenses and increase profits. At the same time, marketing can be carried out in mid-February to prepare for the upcoming increase in sales of this type of product.
   2. Based on Effect of Temperature on Sales: We can conclude that this product is a temperature consuming product. As the temperature rises, its sales will also increase. Companies can consider paying more attention to future temperature forecasts to decide the amount of product launches and the amount of labor. For example, if they see that the temperature will rise next month (not just in January-March) by observing the temperature forecast, they can target this period to sell their products and increase revenue and profits.

## Tools and Technologies:

* SQL, AWS Athena or Excel
* Data visualization tools (Tableau, Power BI, Excel)

## Deliverables:

* A detailed report summarizing data, insights, and recommendations.
* Easy so understand Visualizations to present key insights clearly.
* Insights shared with our stakeholders.

This project provides us with a method to discover business questions from the dataset and to convert them into data questions that can be analyzed using data. We can use tools such as SQL and Excel to analyze the data and draw conclusions. Ultimately, these conclusions will help us predict future trends and determine strategies. This is a very important technique.

# Data Wrangling

## Project Part 1

### Project Description: Data Wrangling for Relations within Public-Trees

Project Title: Data Wrangling Before Analyzing Public-Trees

Objective: The primary goal of this project is to perform comprehensive data wrangling to prepare a robust dataset for Public-Trees. By cleaning, transforming, and consolidating data from various sources, the project aims to enhance the accuracy and usability of customer data for subsequent analysis and reporting.

Background: The City of Vancouver needed to migrate their data to AWS. We were commissioned by them to implement a DAP for them and get answers to business questions from the dataset we selected. Also, we need to make sure the dataset is solid to use for our goals.

Dataset: The data wrangling process will involve only one dataset, here’s what is included in:

* TREE\_ID: ID for trees
* CIVIC\_NUMBER: Civic number for trees
* STD\_STREET: Specific Street name
* GENUS\_NAME: Genus name for trees (Quercus only)
* SPECIES\_NAME: Species name for trees (Acutissima only)
* COMMON\_NAME: Common name for trees (Sawtooth Oak only)
* ON\_STREET\_BLOCK: Street block number
* ON\_STREET: Specific Street area name
* NEIGHBOURHOOD\_NAME: Specific neighbourhood name
* STREET\_SIDE\_NAME: Street side planted condition (Even/Odd)
* HEIGHT\_RANGE\_ID: ID for different Height Range (1, 2, 3, 4, 5, 6, 7)
* HEIGHT\_RANGE: The height range for trees (10-20, 20-30, 30-40, 40-50, 50-60, 60-70, 70-80)
* DIAMETER: Diameters for trees
* DATE\_PLANTED: The date of trees planted
* Geom: JSON format data storing geographic coordinates
* geo\_point\_2d: Easier to use storing geographic coordinates

## Methodology:

1. Data Collection and Data Assessment: I am using data that has been cleaned before. They are ready to be used.
2. Data Cleaning: In this step, I will set some conditions, such as completeness and uniqueness, to ensure the quality of my data. And the “passed” part is prepared for further analysis.

Figure 9

***Data Quality Check Rules***

A screenshot of a computer

AI-generated content may be incorrect.

*Note.* This figure shows the screenshot of the data quality check step in visual ETL created in Glue. Source: AWS Glue

In this step, I checked the completeness and uniqueness for “treeid” column. To see if the completeness is equal or greater than 95% and if Uniqueness is greater than 99%.

Figure 10

***Conditional Router***

A screenshot of a computer

AI-generated content may be incorrect.

*Note.* This figure shows the screenshot of the conditional router step in visual ETL created in Glue. Source: AWS Glue

This step is to separate the quality check output as two files: passed and failed into two different prepared folders.

Figure 11

***Change Schema Before Load***

A screenshot of a computer

AI-generated content may be incorrect.

*Note.* This figure shows the screenshot of the change schema step in visual ETL created in Glue. Source: AWS Glue

The meaning of this step is to drop the useless columns which are for data quality checking for both passed and failed output files. So, it reduces our costs and make the output file looks cleaner.

1. Data Transformation and Consolidation I make sure my data is exported to a reasonable size for further analysis.

Figure 12

***Load Both Passed and Failed into Corresponding Folders***

A screenshot of a computer

AI-generated content may be incorrect.

*Note.* This figure shows the screenshot of the Load Failed/Load Passed step in visual ETL created in Glue. Source: AWS Glue

These two steps are the last step before running the visual ETL. We select the output file format as csv and choose the specified location to output.

## Tools and Technologies:

* AWS Glue
* AWS S3

## Deliverables:

* Make sure the dataset is cleaned to use
* Make sure the output of dataset is successfully and safely transferred to correct place to future analyzing.

This data wrangling project aims to establish a high-quality dataset that enables the City of Vancouver to conduct effective customer analytics, ultimately enhancing marketing strategies, improving customer service, and driving overall business growth.

# Data Quality Control

### Project Description: Data Quality Control Initiative for City of Vancouver

Project Title: Implementation of Data Quality Control Measures at Public-Trees

Objective: The primary objective of this project is to establish a comprehensive Data Quality Control (DQC) framework for the City of Vancouver. This framework will ensure the accuracy, completeness, consistency, and reliability of the organization's data, enhancing decision-making processes and overall business performance.

Background: As the City of Vancouver continues to expand its operations and data sources, issues related to data quality have surfaced, including inaccuracies, duplicate records, and inconsistent formats. Poor data quality can lead to misguided business strategies, inefficiencies, and regulatory compliance risks. This project aims to implement robust data quality control measures to mitigate these issues.

Scope: The project will focus on the following key areas:

* Data Profiling: Analyzing existing datasets to assess quality levels.
* Data Cleansing: Developing processes to correct inaccuracies and eliminate duplicates.
* Data Validation: Implementing validation rules and checks to ensure data integrity.
* Monitoring and Reporting: Establishing ongoing monitoring processes and dashboards to track data quality metrics.

## Methodology:

1. Data Profiling: This step is to prepare the profile for dataset for later cleaning. We profile our dataset in AWS Glue Data Brew.

Figure 13

***Data Profiling***

A screenshot of a computer

AI-generated content may be incorrect.

*Note*. This screenshot shows the profiling result of raw dataset in Data Brew. Source: AWS Glue Data Brew.

1. Data Cleaning: This step is to clean the data before analyzing. I changed the name of the column to clearer format and change the data planted to yyyy-mm-dd format. Also I removed useless part in my dataset during Visual ETL Design, also filter the rows to meet the requirements of data team.

Figure 14

***Data Cleaning***

A screenshot of a computer

AI-generated content may be incorrect.

*Note*. This screenshot shows the cleaning step of raw dataset in Data Brew. Source: AWS Glue Data Brew.

Remove Useless Part in Dataset: In this step, our goal is to remove columns that are not related to our metrics and add filters to obtain the data we need. This makes our dataset look neater and more straightforward.

Figure 15

***Change Schema of Summarization 1***

A screenshot of a computer

AI-generated content may be incorrect.

*Note.* This screenshot shows the change schema step in AWS Glue Visual ETL. Source: AWS Glue

I use “Change Schema” to drop some useless columns which I don’t want it to appear in my summarization. And I change the data type of date planted to timetable format. Finally, I only keep 5 columns here.

Figure 16

***Filter of Summarization 1***

A screenshot of a computer

AI-generated content may be incorrect.

*Note.* This screenshot shows the filter step in AWS Glue Visual ETL. Source: AWS Glue

I use filter to filter my rows and keep those the diameter is equal or greater than 10. Now I only have 200 rows cut from 500 rows.

1. Validation Rules and Procedures: This step is to ensure that our output data is stored in a safe and appropriate location after analysis. In this case, I set up two folders, user and system, to store two different storage locations that meet customer and system requirements. I set the target location for both the user and system for loading the output.

Figure 17

***Set Target Location***

A screenshot of a computer

AI-generated content may be incorrect.

*Note.* This screenshot shows the Load step in AWS Glue Visual ETL. Source: AWS Glue

Figure 18

***User Output of Summarization 1***

A screenshot of a computer

AI-generated content may be incorrect.

*Note.* This screenshot shows the user output of summarization 1 in AWS S3. Source: AWS S3.

Here’s the output after summarization. It is stored as a user-friendly format in the user-folder .

Figure 19

***System Output of Summarization 1***

A screenshot of a computer

AI-generated content may be incorrect.

*Note.* This screenshot shows the system output of summarization 1 in AWS S3. Source: AWS S3

Here’s the output of summarization 1. It is stored as a system-friendly format in the system folder.

1. Monitoring and Reporting:

## Cleaning: Week 3 Class Activity

Figure 20

***Poor Quality Dataset***A screenshot of a computer

AI-generated content may be incorrect.

*Note.* This screenshot shows the poor-quality dataset activity in week 3. Source: By UCW

This is the class activity we did in week 3. We can find many problems in this poor-quality dataset, such as missing value, typo issue, duplicated issue, etc. By discovering the problems contained in each column, we can effectively exercise the ability to prepare cleaned datasets and prepare for possible analytical work in the future.

## Profiling: Week 3&4 Activity

Figure 21

***Week 3 Activity Screenshot*** ***A screenshot of a computer

AI-generated content may be incorrect.***

Note. This is the screenshot of profiling week 3 activity.

In this week, We learned how to create a profile for our dataset. The profile allows us to clearly see the current status of the dataset after summarization, such as whether we have missing values. We can also see the distribution of some numerical values.

## Tools and Technologies:

* Excel
* AWS Glue Data Brew
* AWS Glue
* AWS S3

Deliverables:

* A comprehensive Data Quality Control plan detailing processes, metrics, and responsibilities.
* Documentation of data quality metrics and KPIs being tracked.
* Cleaned and validated datasets ready for analysis and reporting.
* A monitoring dashboard that visualizes data quality metrics in real-time.

## Timeline:

* Expected completion of the project: From Week 3 to Week 12.

This Data Quality Control initiative aims to empower City of Vancouver to enhance its data integrity and reliability, resulting in improved decision-making, operational efficiency, and compliance with regulatory requirements.

# Course Completion Badge

Share your course completion badge in your portfolio. You can claim it using a module named “Badges and completion certificates” in your AWS Academy CF course.