Part 1: Descriptive Analysis

Descriptive Question: "What is the average speed limit of bikeways in Vancouver by type of bikeway?"

Step 1: Data Ingestion

The objective of this step is to upload the dataset “bikeways2.csv” to AWS S3 Bucket. By creating a file in S3 bucket. The dataset is stored in the raw bucket as unprocessed data.

A screenshot of a computer

Description automatically generated

Step 2: Data Profiling

The objective this step is to understand the characteristics and quality of the dataset. In AWS console search for AWS Glue DataBrew and open it. Then create a project with suitable title and choose the raw data from S3 bucket.

Once the data is been uploaded then run the job. Once the running of job profiling is done open data cleaning.

This profiling helps us to know the missing values and the data quality.

A computer screen shot of a computer screen

Description automatically generated

Step 3: Data Cleaning

The objective of this step is to clean data to ensure accuracy and good quality for analysis.

Once you open the project click on add steps to edit the dataset by adding the missing values to the data set and changing the format of the dataset and deleting the columns.

A screenshot of a computer

Description automatically generated

Once after cleaning the dataset and need to run the dataset to have good quality dataset to get the accurate results

A computer screen with a message

Description automatically generated

Once cleaning steps are defined, the dataset needs to be stored back in the S3 bucket in 2 segments one is “System” and the other is “user”. This cleaning steps ensures that inconsistencies do not affect the analysis.

A computer screen with a white background

Description automatically generated with medium confidence

Step 4: Data Pipeline Design

The Objective of Data Pipeline Design is to design an automated pipeline for the entire data flow.

To create a data pipeline need to open AWS Glue in the AWS console, and create a new visual ETL. And import the data and filter the data and find out the average speed limit of bikeways in Vancouver by type of bikeway by calculating the total number of bike ways in Vancouver and the speed limit of the each type of bikeways and the overall average speed limit of bikeways in Vancouver and save the output in 2 different file as “System” & “User”.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

After saving the output the dataset will be reflecting in the S3 bucket and the Implementation Screenshots are added below.

A computer screen with a white background

Description automatically generated

A screenshot of a computer

Description automatically generated

Part 2: Exploratory Analysis

Exploratory Questions: “How do speed limits impact the frequency of snow removal on different types of bikeways?”

Step 1: Data Ingestion

The objective of this step is to upload the raw dataset to the S3 Bucket and use to analysis the Exploratory analysis.

A screenshot of a computer

Description automatically generated

Created a Raw file in S3 bucket and uploaded the dataset “bikeways2.csv” and kept it ready for analysis.

Step 2: Data Profiling

After storing the dataset in S3 bucket now it’s turn to the clean the data, so before cleaning the dataset need to quality of the dataset because we have the poor dataset now and need to know the quality and where to fix the dataset.

For that data profiling should be done open ASW DataBrew and update the dataset and know the quality of the data and move the dataset to the cleaning platform in AWS .

A screenshot of a computer

Description automatically generated

While profiling it helps in understanding the frequency and geographic areas affected by snow removal.

Step 3: Data Cleaning

A screenshot of a computer

Description automatically generated

The objective of the data cleaning is to fill in the missing values and convert the poor-quality data to good-quality data for accurate analysis. In this handle missing values by filling in the most frequent number for the speed limit and changing the data format for accurate results.

A computer screen with a white background

Description automatically generated

After cleaning the data saving them for easy access in transfer bucket in AWS under 2 different folders “system” & “User”.

A computer screen with a white background

Description automatically generated

Step 4: Data Pipeline Design

So after ingestion profiling and cleaning the pipeline design helps us to understand the flow of the dataset and clearly explain the merging reason and the get the detailed analysis fo the exploratory analysis.

So first create a workflow and inject the data to the pipeline after that clean the data by dropping the unwanted row and columns in the cell. After that calculating the snow-removed bike ways and understood the exploratory analysisand save the output in 2 different file as “System” & “User”.

A screenshot of a computer

Description automatically generated

After saving the output the dataset will be reflecting in the S3 bucket and the Implementation Screenshots are added below.

A screenshot of a computer

Description automatically generated

A computer screen with a number of text

Description automatically generated with medium confidence

Data Visualization:

This Chart compares the speed limit and snow removal for each street. And it can be used to analysis descriptive and exploratory analysis.

A screenshot of a computer

Description automatically generated