Pre-Owned Cars on the US market

Project Plan

Version 1.0

Revision History

| **Date** | **Version** | **Description** | **Author** |
| --- | --- | --- | --- |
| 11/05/2021 | 0.1 | Initial Draft -Draft 1 | Purnima Bhukya |
| 11/08/2021 | 0.2 | Draft 2 | Nivedha Balakrishnan |
| 11/14/2021 | 0.3 | Draft 3 | Michele Assouad |
| 11/16/2021 | 0.4 | Draft 4 | Praveen Kuruvangi Parameshwara |
| 11/17/2021 | 1.0 | Baseline | Praveen Kuruvnagi Parameshwara |

Table of Contents

1. Introduction 4

1.1 Purpose of this document 4

1.2 Intended Audience 4

1.3 Scope 4

1.4 Definitions and acronyms 4

1.4.1 Definitions 4

1.4.2 Acronyms and abbreviations 5

1.5 References 5

2. Background and Objectives 5

3. Organization 5

*3.1* Project group 5

3.2 Customer 6

4. Development process 6

*5.* Deliverables 6

6. Project risks 6

7. Communication 7

7.1 Canvas 7

7.2 Git 7

8. Project plan 7

8.1 Time schedule 7

8.1.1 Remarks 7

8.2 Test plan 9

8.2.1 Testing Remarks 10

9. References 10

# Introduction

## Purpose of this document

The purpose of this document is to provide a detailed project description of the application called ‘Pre-owned cars in the US market’, which is designed to help people understand the factors that affect the price of a used car and guide their decision when buying a used car. This document includes details about the data gathered between 1915 and 2021 for pre-owned cars postings on the website Craigslist, insights from the data, deliverables, time plans and financial plans.

## Intended Audience

This document shall be used in all phases of the project as a guideline. Intended audiences of this project are all project stakeholders:

* project leader: Praveen Kuruvangi Parameshwara
* team members: Michele Assouad, Nivedha Balakrishnan, Purnima Bhukya

## Scope

Our project aims to analyze a dataset that is a collection of entries for all the used cars put on sale on the website “Craigslist”, over the United States’ territory, between 1915 and 2021 and provide graphical dashboard that detail the factors that have a major role in setting the price of the pre-owned car.

## Definitions and acronyms

### Definitions

| **Keyword** | **Definitions** |
| --- | --- |
| Pre-owned Cars | The name of the project |
| Project Supervisor | Andrew H. Bond |
| Project Leader | Praveen Kuruvangi Parameshwara |
| Team Member | Michele Assouad  Nivedha Balakrishnan  Purnima Bhukya |
| Milestone | Aug -21 to Dec -21 |
| Git | Version control system that will be used in this project |
| Scrum | An iterative and incremental agile software development method for managing software projects and product or application development |
| Tableau | Data analysis tool to build graphs, charts and Dashboard |
| Python | Data analysis tool to view granularity of data and to build charts. |
| Redshift | Aws Relational database to store the data. |
| Web Designer | Purnima Bhukya |
| Amazon Route 53 | cloud Domain Name System (DNS) web service |

### Acronyms and abbreviations

| **Acronym or**  **abbreviation** | **Definitions** |
| --- | --- |
| AWS | Amazon Web Services |
| S3 | AWS Simple Storage Service |
| RDS | AWS Relational Database Service |
| VPN | Virtual Private Network |
| IAM | Identity and Access Management |
| ARN | Amazon Resource Names |

## References

1. https://www.kaggle.com/ananaymital/us-used-cars-dataset?select=used\_cars\_data.csv
2. https://www.kaggle.com/austinreese/craigslist-carstrucks-data
3. https://www.kaggle.com/tsaustin/us-used-car-sales-data
4. https://simplemaps.com/data/us-zips
5. https://www.consumerreports.org/buying-a-car/car-buying-guide/
6. https://cars.usnews.com/cars-trucks/how-to-buy-a-used-car

# Background and Objectives

The pre-owned cars have greater advantage in the market because of low cost, less depreciation and low insurance. In the past two years because of COVID situation, more support towards work from home, travel restriction and quarantine, the number of cars that were available in the market were more in numbers. Beginning of this year, the situation has improved and demands on the pre-owned cars have increased. The demand has been increasing drastically since Feb-2021. We have observed this trend and decided to explore more on the pre-owned cars. We wanted to understand the main factors that customers focus on to buy cars such as year, make, model, color, fuel and others.

# Architecture & High Level Design

1. Initially, the denormalized dataset consisted of 3 millions rows and 66 columns. The data is then cleaned and transformed using Python and MySQL Workbench.
2. After normalization, the data is converted into 8 tables and loaded to the Amazon S3 bucket as csv files. Each csv file represents each table in the data model.
3. In this project, we exported the data from S3 bucket using two ways.
4. First, the data from S3 bucket is transferred to Amazon RDS database instance using AWS Lambda function. Whenever the csv file is uploaded into the specified S3 bucket, lambda

function triggers in such a way that the data from S3 is transferred to an RDS database instance. The database in RDS is then connected to the MySQL workbench where we can query and modify the data.

1. Second, the data is exported from S3 bucket to Amazon Redshift cluster using Redshift Query Editor.
2. The database from both AWS RDS and AWS Redshift is connected to tableau and python for visualization.
3. The Web application is developed in AWS using the Route 53 domain.

# Organization

## Project group

| **Name** | **Initials** | **Responsibility (roles)** |
| --- | --- | --- |
| Praveen Kuruvangi | PK | Team lead/Data modeling and Visualization |
| Nivedha Balakrishnan | NB | Team member/Data Architecture |
| Michele Assouad | MA | Team member/Data Cleaning |
| Purnima Bhukya | PB | Team member/Visualization and Web Development |

## Customer

The target customers are listed below:

* Data Analysts
* Used Cars Buyers

# Development process

The project will use Python language and Jupyter Notebooks to store and share the data cleaning code, MySQL Workbench and MySQL to build the data model, AWS S3 as the data lake to load and store the data on the cloud, AWS RDS to store the data on the cloud in a relational database by triggering a Lambda function using AWS Lambda and connecting AWS RDS to MySQL Workbench to query the data, AWS Redshift as a data warehouse on the cloud to store and query the data, Tableau and Jupyter Notebook to create dashboards and graphs to visualize the data, AWS Route 53 to create and host a web application related to the dataset and the visualization.

# Deliverables

| **To** | **Output** | **Planned Week** | **Promised week** | **Late +/-** | **Delivered week** | **Notes** |
| --- | --- | --- | --- | --- | --- | --- |
| Data Extraction | Data was downloaded from Kaggle | Sep 13 | Sep 13 | no | Sep 13 |  |
| ​​Data cleaning and wrangling | Data was cleaned using python | Oct 17 | Oct 17 | yes | Oct 24 |  |
| Data Modeling | Schema and ER Diagram on MySQL Workbench | Oct 24 | Oct 24 | no | Oct 24 |  |
| Data Loading to Data Lake | Data was loaded to an AWS S3 bucket | Oct 31 | Oct 31 | no | Oct 31 |  |
| Data loading to Data Warehouse | Data was loaded to an AWS Redshift cluster | Oct 31 | Oct 31 | no | Oct 31 |  |
| Visualization and analytics | Data was analysed using Python and Tableau | Nov 7 | Nov 7 | no | Nov 7 |  |
| Web Application | Project info and visuals are on the web by connecting AWS Route 53 with S3 | Nov 14 | Nov 14 | no | Nov 14 |  |
| Project Presentation |  | Nov 14 | Nov 14 | no | Nov 14 |  |
| Project Report |  | Nov 14 | Nov 14 | no | Nov 14 |  |
|  |  |  |  |  |  |  |

# Project risks

| **Possibility** | **Risk** | **Preventive action** |
| --- | --- | --- |
| Cost risk for using s3 and cluster on AWS Redshift | Storage and resource | Identify unnecessary data and load only required data. Take an action to pause the clusters when not in use. |
| Additionally information requirement to initiate deeper analysis | Data might not give insight. | State information was not available as part of the data. So additional information “location” has been added to identify the state and country. |
| High volume of data. | More space | Image columns and multi-attributes columns that are not part of relational analysis have been moved out to reduce the space overhead. |

# Communication

Below effective communication channels are in place.

1. ZOOM meeting on weekly basis
2. Offline updates on the documents through Shared Drive.
3. Status updates through emails.

## Collaboration

The project was divided into multiple tasks and distributed among the team members. Weekly meetings to understand the issues and concerns. Work together to resolve the issue efficiently and plan the next steps.

## Git

All source code and finished documentation will be uploaded to Github repository. ..

Repository URL: <https://github.com/PurnimaSJSU/Big-Data-Project>

# Project plan

## Time schedule

| **Id** | **Milestone**  **Description** | **Responsible Dept./Initials** | **Finished week** |  |  |  | **Metr.** | **Rem.** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Plan** | **Forecast** |  | **Actual** |  |  |
|  |  |  |  | **Week** | **+/-** |  |  |  |
| 1 | Data Extraction | Everyone | Sep 13 |  |  | Sep 13 |  |  |
| 2 | Data Cleaning | Michele | Oct 17 |  |  | Oct 24 |  |  |
| 3 | Data Modeling | Praveen | Oct 24 |  |  | Oct 24 |  |  |
| 4 | Data Loading to AWS S3 | Nivedha | Oct 24 |  |  | Oct 31 |  |  |
| 5 | Data loading to AWS Redshift | Nivedha | Oct 31 |  |  | Oct 31 |  |  |
| 6. | Data loading to AWS RDS | Nivedha | Oct 24 |  |  | Oct 31 |  |  |
| 7. | Visualization in Tableau | Praveen | Nov 7 |  |  | Nov 7 |  |  |
| 8. | Visualization in Python | Purnima | Nov 7 |  |  | Nov 7 |  |  |
| 9. | Web page application using AWS Route 53 | Purnima/Michele | Nov 14 |  |  | Nov 14 |  |  |

### Remarks

| **Remark Id** | **Description** |
| --- | --- |
|  |  |

## Test plan

| **Test No.** | 001 | **Phase:** | 1 | **Author:** | Michele, Nivedha | Date: 11/15/2021 |
| --- | --- | --- | --- | --- | --- | --- |
| **Test Category:** | | **Extract the data and load into S3 and then to databases** | | |  |  |
| **Software Product:** | | CSV, Redshift | | | |  |
| **Test Title:** | | Source and destination row count. | | | | |
| **Test Purpose:** | | Source and destination row count should match | | | | |
| **Test Setup:** | | Configure AWS lamda functions | | | | |
| **Prerequisites:** | | CSV files | | | | |
| **Procedure:** | | Separate the CSV files in python after the cleaning process and load the data to S3 | | | | |
| **Checks:** | | Verify the field values and its datatypes | | | | |
| **Expected Results:** | | Rowcount and data should load to the respective field without any errors. | | | | |
| **Result:** | | Source and destination row count matched. | | | | |
| **Reason for Failure:** | | No failure | | | | |
| **Remarks:** | |  | | | | |

| **Test No.** | 002 | **Phase:** | 1 | **Author:** | Praveen, Purnima | Date: 11/15/2021 |
| --- | --- | --- | --- | --- | --- | --- |
| **Test Category:** | | Correct data display on grap**h** | | |  |  |
| **Software Product:** | | Tableau, Python | | | |  |
| **Test Title:** | | Correct data to display on graph | | | | |
| **Test Purpose:** | | Correct data should display on graph | | | | |
| **Test Setup:** | | Rechecks were done to confirm the data displayed to match the actual data source. | | | | |
| **Prerequisites:** | | Data should be loaded in s3 | | | | |
| **Procedure:** | | Recheck were done to confirm that displayed data is matching with actual data source | | | | |
| **Checks:** | | Recheck were done to confirm that displayed data is matching with actual data source | | | | |
| **Expected Results:** | | Graph data should match to the s3 loaded data. | | | | |
| **Result:** | | Graph data is matching s3 loaded data. | | | | |
| **Reason for Failure:** | | No Failures | | | | |
| **Remarks:** | |  | | | | |

### Testing Remarks

| **Remark Id** | **Description** |
| --- | --- |
| 1 | Verification of primary and foreign key relationships |
| 2. | Validation of data and its data type |
| 3 | Record count check on both fact and dimension tables |
| 4 | Data profiling, Data mining and Data analysis |

# References

* <http://sjsudatabusters.com.s3-website-us-west-1.amazonaws.com/>
* <https://github.com/PurnimaSJSU/Big-Data-Project>
* <https://www.kaggle.com/ananaymital/us-used-cars-dataset?select=used_cars_data.csv>
* <https://www.kaggle.com/austinreese/craigslist-carstrucks-data>
* <https://www.kaggle.com/tsaustin/us-used-car-sales-data>
* https://simplemaps.com/data/us-zips
* https://www.consumerreports.org/buying-a-car/car-buying-guide/
* <https://cars.usnews.com/cars-trucks/how-to-buy-a-used-car>