Data 228

Yelp Insights

Project Plan

Version 1.0

Abraca Data (Group 7)

Table of Contents

1. Introduction

1.1 Purpose of this document

This document provides a detailed project plan of the application - Yelp Insights.

1.2 Intended Audience

- New investors in restaurant business
- Existing restaurant owners
- Customers visiting restaurants

1.3 Scope

YELP and US Census Bureau heterogeneous datasets will be combined to find impact of average income of households in counties on restaurant businesses in those counties in US. Using the extended attributes of the dataset we can find insights of the comparison between counties having the same household income vs the same restaurant price/cost ($$), Restaurant Reviews and their patterns like most popular restaurants, most popular cuisine etc.

1.4 Definitions and acronyms

1.4.1 Definitions

Keyword	Definitions
Project Name	Yelp Insights
Project Leader	Vani Kancherlapalli
Team Member	Chidroop Sagar, Manisha Paliwal
Milestone	Sep 2021 - Nov 2021
Git	https://github.com/vanikancherlapalli/Abraca-Data-228
Scrum	An iterative and incremental agile software development method for managing software projects and product or application development
Scrum sprint	Weekly
Scrum master	Manisha Paliwal
Product owner	Chidroop Sagar

1.5 References for data source

- https://www.yelp.com/dataset
- https://data.census.gov/cedsci/table?q=Income%20%28Households,%20Families,%20Individuals%29&g=0400000US01%240500000,02%240500000,04%240500000&tid=ACSST1Y2019.S1901&hidePreview=true
- https://www.unitedstateszipcodes.org/

2. Background and Objectives

Yelp Inc. is an online portal which provides crowd-sourced reviews for businesses and extends other services like restaurant table reservation. During the COVID-19 pandemic situation, ordering food online was the preferred source for people due to shelter in place orders issued by the Government and their safety. Yelp being the most popular review portal provided an edge to make informed decisions for the people during the Pandemic. This can be an opportunity for business owners to expand their sales by keeping their Yelp profile competitive with increase in dependency of people on reviews and ratings.
Our project focuses on analyzing the current trends of correlation between income and restaurants business, impact of business attributes, customer's reviews on business ratings and their price range
In this project we will be utilizing AWS-provided services and third-party software, Python, MySQL and Tableau to model, wrangle and analyze the data. The project findings will help investors to figure out the feasibility of a new venture and guide existing owners how to upscale their business margins

3. Architecture & High-Level Design

Project utilizes Amazon web services, python, and Tableau as below:

- **Amazon S3(Simple Storage Services)** to store raw data sets
- **AWS Glue and Python libraries** NumPy, Pandas and JSON were sued to perform ETL processes like data cleansing, filtering, removing unwanted fields and fattening of Json into tabular format
- **AWS Redshift** is used as relational database to store cleansed data
- **Tableau** is used to perform analytical processes and visualization.
- **Website** is created for end users to view the work

4. Organization

4.1 Team Members

Student Name	Student ID	Role
Vani Kancherlapalli	014702207	Team Lead
Chidroop Sagar	015926846	Team Member
Manisha Paliwal	015935374	Team Member

4.2 Customer

The target customers are listed below:

- New investors in restaurant business
- Existing restaurant owners
- Customers visiting restaurants
- Business Analyst

5. Development process

- Analyze the raw dataset YELP Business and Review Files, US Census Average household income and zip code dataset
- Perform data wrangling as below:
  I. Yelp Business (JSON file) was uploaded in S3 bucket. Using Glue Job nested JSON structure was flattened into tabular columns, cleansed, filtered and required fields were loaded into AWS Redshift Cluster
  II. Yelp Review (JSON file) was uploaded in S3 bucket. Using COPY command filters and JSON schema file data was flattened into columnar data and loaded into AWS Redshift cluster

III. Yelp Business Ambience Attributes (nested string format) was fetched from Yelp business file and converted into JSON format using python. After conversion processed file was uploaded in S3 and AWS Redshift Cluster
IV. US Census Household Income (csv file) was uploaded in S3 bucket. Using Glue Job data was cleanse, filtered and required fields were loaded into AWS Redshift Cluster
V. Zip Code (csv file) was converted to text pipe delimited file to avoid truncation of leading zeros in ZIP Code using python and then uploaded in S3 bucket. Further copied to AWS Redshift using features of COPY command

- Create visualization in Tableau and publish tableau story in tableau public
- Create website and host it using AWS S3 website hosting

6. Deliverables

- Project Plan
- Project Presentation
- Project Report [Technical Document]
- Project Source Code shared on GitHub

7. Project risks

Possibility	Risk	Preventive action
Yelp Dataset has a size of 4.5 GB	This might be trimmed to adhere to AWS cost	ETL was performed such that project relevant dataset was retained post ETL process
Consistency in the three datasets used and merge those with common key	It could lead to data inconsistency and compromised data integrity	Relationship was created between datasets using Primary and foreign keys. Data Modelling was performed
Cost risk for using S3 and Cluster on AWS	It can increase if not used appropriately	we have paused our cluster when it was not in used and used optimized queries for data analysis
Technical glitches during integration of AWS redshift and tableau	It can result in missing timelines	Proactively performed POCs to check feasibility of project design and took corrective measures to integrate tools

8. Communication

8.1 Collaboration

All team members were connected through Zoom call weekly and followed agile methodology to execute the project

8.2 Git

All source code and finished documentation uploaded to GitHub repository

Repository URL: https://github.com/vanikancherlapalli/Abraca-Data-228

9. Project plan

9.1 Time schedule

To	Output	Planned Week	Delivered Week	Late +/-
Project Design	Design the architecture of project, tools, and technologies to be used	6th Sep	15th Sep	No
Data Analysis	Data was downloaded from Source and analyzed for ETL Scope	15thth Sep	24th Sep	No
ETL	Data wrangling and related coding in Glue, python, S3	27th Sep	11th Oct	No
ETL Testing	Testing source and target data	15th Oct	27th Oct	
Visualization	Data was analyses using tableau	20th Oct	1st Nov	No
Visualization Testing	Graphs were tested for correctness	1st Nov	7th Nov	No
Website Development	Website design and coding	1st Nov	10th Nov	No
Website Testing	Testing website across browsers	8th Nov	14th Nov	No

Website Deployment	Host website using AWS S3 Web hosting	15th Nov	16th Nov	No
Documentation	Project work documentation	24th Sep	20th Nov	No

9.2 Test plan

Test No.	001	**Phase:**	1	**Tester:**	Vani/Manisha	Oct 2021
Test Category:		**ETL Testing**				
Software Product:		AWS Glue, Python, AWS S3, Redshift				
Test Title:		Source and destination data count and validation				
Test Purpose:		Source and destination data count and validation				
Test Setup:		Manual Testing was performed by querying source and target				
Prerequisites:		AWS Glue job and source data				
Procedure:		Source Dataset uploaded in AWS S3 was compared with Target Data post ETL using SQL queries				
Expected Results:		Target dataset should be loaded as expected				
Result:		Target dataset was loaded as expected				
Reason for Failure:		No failure				
Remarks:		Test Results Passed				

Test No.	002	Phase:	1	Author:	Vani, Chidroop, Manisha	Nov 2021
Test Category:	**Visualization Test**					
Software Product:	Tableau					
Test Title:	correct data display on graph					
Test Purpose:	correct data should display on graph					
Test Setup:	rechecks were done to confirm that displayed data is matching with actual data source					
Prerequisites:	data should be loaded in AWS Redshift					
Procedure:	rechecks were done to confirm that displayed data is matching with actual data source					
Checks:	rechecks were done to confirm that displayed data is matching with actual data source					
Expected Results:	graph data should match query results from AWS Redshift					
Result:	graph data matched query results from AWS Redshift					
Reason for Failure:	No failure					
Remarks:	Test Results Passed					

Test No.	003	Phase:	1	Author:	Chidroop Sagar	Date: Nov 2021
Test Category:	**Website testing**					
Software Product:	AWS S3 Web hosting and multiple screens					
Test Title:	Website hosted successfully and is compatible with multiple screens					
Test Purpose:	Website hosted successfully and is compatible with multiple screens					
Test Setup:	Browsing website using different browsers					
Prerequisites:	Website source code created and hosted using AWS S3					
Procedure:	Website link tested for accessibility and compatibility over multiple screens					
Expected Results:	Website link should be accessible and compatible					
Result:	Website link was accessible using different browsers					

Reason for Failure:	No failure
Remarks:	Compatibility on Google chrome is the best. Needs more improvement in internet explorer

9.3 Project Roles and Responsibilities

Project Work	Ownership
Dataset Proposal	
YELP	Chidroop Sagar
US Census bureau Household Income	Manisha Paliwal
Zip Code Database	Manisha Paliwal
Project Architecture	
Design Plan	Vani Kancherlapalli
Database Data Model	Manisha Paliwal
Source File- AWS S3 Upload	
Yelp Business	Vani Kancherlapalli
Yelp Review	Vani Kancherlapalli
US Census Income Dataset	Manisha Paliwal
Zip Code Dataset	Manisha Paliwal
ETL and related testing/validation	
AWS Glue - Redshift	
Yelp Business	Vani Kancherlapalli
Yelp Review	Vani Kancherlapalli
US Census Income Dataset	Manisha Paliwal
Python -S3- Redshift	
Yelp Business Attribute Ambience	Manisha Paliwal
Zip Code Dataset	Manisha Paliwal
Visualization	
Word Cloud of Cuisines	Manisha Paliwal
Top Ten Restaurants by Count	Manisha Paliwal
Average Rating of Top Ten Restaurants	Manisha Paliwal
Relation between Average Household Income of County and Number of Restaurant	Manisha Paliwal
Distribution of Restaurant Price Range based on Average Household Income	Manisha Paliwal
Number of Reviews vs Restaurant Rating	Manisha Paliwal
Relation between Average Length of Review and Review Rating	Manisha Paliwal
Comparison between Ambience vs Estimated Average Income Range	Chidroop Sagar
Effect of Food Categories on Ratings	Chidroop Sagar
Effect of Food Categories on Ratings Review By Years	Chidroop Sagar
Restaurant Price ($) and stars	Vani Kancherlapalli

Restaurant Attributes Compare	Vani Kancherlapalli
Restaurant Attributes Compare of restaurants with 2 stars	Vani Kancherlapalli
Business review counts with more analysis on stars count	Vani Kancherlapalli
Website Hosting in AWS S3	
Website Design	Manisha Paliwal
Website Content	Vani Kancherlapalli, Chidroop Sagar, Manisha Paliwal
Website Testing	Vani Kancherlapalli, Chidroop Sagar, Manisha Paliwal
Website Hosting in AWS S3	Chidroop Sagar

10. References

Website:

- https://websiteyelpinsights.s3.us-west-1.amazonaws.com/YelpInsights/Home.html

Source:

- https://data.census.gov/cedsci/table?q=Income%20%28Households,%20Families,%20Individuals%29&g=0400000US01%240500000,02%240500000,04%240500000&tid=ACSST1Y2019.S1901&hidePreview=true
- https://www.yelp.com/dataset
- https://www.unitedstateszipcodes.org/

GitHub:

- https://github.com/vanikancherlapalli/Abraca-Data-228

Architecture Diagram:

- https://app.diagrams.net/

Tableau Dashboard:

- https://public.tableau.com/app/profile/manisha.paliwal/viz/YelpInsightsByAbracaData/YelpInsights?publish=yes
- https://public.tableau.com/app/profile/vani.k4703/viz/YelpInsightsByAbracaData_Vani-1/YelpInsights?publish=yes
- https://public.tableau.com/app/profile/chidroop.sagar1502/viz/YelpInsightsbyAbracaData/AbracaDataVisualisationCs?publish=yes

Images:

- https://www.census.gov/
- https://www.yelp.com/dataset
- https://www.dictionary.com/e/zip-code/
- https://www.clipartmax.com/middle/m2H7K9G6Z5K9i8Z5_future-scope-clipart-man-with-binoculars-png/

Website Development Reference:

- Main Image
- Yelp Image
- US Census Image
- Zip Code Image