King Saud University
College of Computer and Information Sciences
Department of Information Technology
IT 362 – Data Science
2st Semester 1446 H

Analyzing Freelancing Trends and Sustainability

Logbook

	Students Name	ID
Section #66850 Group #6	Layan Alfawzan	444200793
	Rose Mady	444200107
	Aljawharah Alwabel	444200750
	Arwa Almutairi	444201055

Supervised by: Dr.Reem Alqifari

Phase	Task	Date	Decisions & Rationale	Challenges & Solutions	Implementation
1: Data Collection Research and Assessment	Web Scraping "Freelancer.com/freelancers/"	29-Jan-2025	Scraping 1,000 samples from the website	Inconsistent data retrieving which is going to be solved by dealing directly with scraped csv dataset file.	
	Writing Dataset Overview	02-Feb-2025	Analyzed dataset structure, checked for missing values, and assessed data quality.	The dataset contains 65 missing values. Also, 104 duplicate records were identified in 'Freelancer Name' and will be addressed by retaining only unique entries.	
	Handling Missing Values	20-Feb-2025	Filled missing "Skills" with "Unknown" and missing "Reviews" with the median. Reviews were set to 0 where the rating was 0.	Ensuring logical consistency between rating and reviews while preventing unnecessary data loss.	Used df.fillna() and logical conditions to handle missing values.
	Removing Duplicates	20-Feb-2025	Removed duplicate freelancer names, keeping only the first occurrence.	Prevented overrepresentation of duplicate freelancers.	Used df.drop_duplicates(inplace=True).
	Converting Data Types	1-Mar-2025	Extracted numeric values from "Hourly Rate" and converted them to numerical format.	Standardizing currency formats for accurate analysis.	Used df['Hourly Rate'] = pd.to_numeric(df['Hourly Rate'].str.replace(r'[^0-9]', ", regex=True)).
	Text Processing for Bio Column	3-Mar-2025	Tokenized text, removed non-alphabetic characters, stopwords, and duplicate words, and applied lemmatization.	Ensured relevance by retaining only meaningful words while handling missing bio entries.	Used NLTK’s word_tokenize, stopwords, and WordNetLemmatizer.
2:Data Collection, Processing, Cleaning, and Exploratory Data Analysis (EDA)	EDA: Non-Graphical univariate	1-Mar-2025	Computed summary statistics for key numerical columns.	Skewed distributions required additional insights for correct interpretation.	Used df.describe() and df.median().
	EDA: Non-Graphical multivariate	1-Mar-2025	Analyzed relationships between rating, reviews, hourly rate, and total earnings.	The weak correlation between hourly rate and earnings required deeper insights.	Used df.corr() and seaborn heatmaps.
	EDA:Graphical univariate	3-Mar-2025	Visualized distributions of Hourly Rate, Reviews, Rating, and Skills Count to understand spread and identify outliers.	Skewed distributions in hourly rates and reviews required careful interpretation.	Used seaborn histplot(), countplot(), choropleth(), barplot(), boxplot(), wordclod using imshow()

	EDA:Graphical multivariate	3-Mar-2025	Plotted scatter plots and correlation matrix to explore relationships between earnings, rating, reviews, and skills.	Outliers impacted visualization clarity; solved by adjusting axis limits and point transparency.	Used seaborn scatterplot(), boxplot(), heatmap().
	Outlier handling	14-Mar-2025	Used IQR method to detect outliers to reduce skewness.	Extreme values affected distributions; solved by removing specific extreme cases.	By deleting the value
	Normalize numerical features	18-Mar-2025	apply Min-Max normalization on key numeric columns to bring them to a common scale (0-1).	Columns had different scales, which is solved by normalizing each using its own min and max values.	MinMaxScaler()
	Encode categorical column	18-Mar-2025	Label Encoding for 'Location', and counting number of skills per freelancer for 'Skills' column .	Challenge with varying formats, which was addressed by using label encoding for 'Location' and extracting skill count from the 'Skills' column.	LabelEncoder, .split(',')
3: Modelling and Communication	Train Baseline Linear Regression Model	9-Apr-2025	Building initial model to establish a reference point for future improvements.	Ensured fair evaluation by splitting data and testing model accuracy using key metrics.	Used train_test_split, LinearRegression, mean_squared_error, and r2_score.
	Train Random Forest Model	9-Apr-2025	Used an ensemble model to improve prediction accuracy over the baseline.	Managed model complexity and performance by evaluating using test data and key metrics.	Used train_test_split, RandomForestRegressor, predict, r2_score, and mean_squared_error.
	Train and evaluate XGBoost model	9Apr-2025	Chose XGBoost for its high accuracy and performance with regression tasks.	Managed model complexity and performance by evaluating using test data and key metrics.	Used train_test_split, XGBRegressor, predict, r2_score, and mean_squared_error.
	Determining optimal number of clusters	30-Mar-2025	Used Elbow Method (WCSS) and Silhouette Score to find the optimal number of clusters.	Challenge with identifying the exact elbow point and ensuring well-defined clusters, solved by Combining elbow method and silhouette score for improved evaluation.	KMeans, KneeLocator, silhouette_score
	K-Means Clustering and Evaluation (Elbow + Silhouette)	30-Mar-2025	Chose k-values for clustering, using silhouette scores and visualization tools.	selecting the optimal k for K-means clustering and assessing cluster quality, K-means was applied for k=2 to k=4, evaluated using silhouette scores, and visualized with Yellowbrick for quality assessment.	KMeans, silhouette_score, fit_predict, plot, SilhouetteVisualizer, inertia_, random_state, n_init.
	Visualized K-means clusters	30-Mar-2025	Used scatter plot to visualize the K-means clustering results.	Challenged distinguishing clusters in high-dimensional data. Solution: Focused on two features for easier visualization of clusters.	sns.scatterplot
	Find Optimal Number of Clusters for Hierarchical Clustering	1-Apr-2025	Used silhouette scores for k=2 to k=10 to determine the optimal number of clusters,	Challenge with determining the best k value for clustering; solved by silhouette scores for k=2 to k=10 and	AgglomerativeClustering, silhouette_score

				evaluate to choose the optimal k value.	
	Clustering and Visualization	1-Apr-2025	Visualized clusters and analyzed reviews vs. total earnings.	Challenge with effective visualization to clusters in the data, scatterplot and box plots are used to display the distribution of total earnings across clusters.	scatterplot, boxplot
	Evaluation of Clustering Quality	1-Apr-2025	Evaluated clustering quality using silhouette score and statistical validation.	Verifying the quality and consistency of clusters; Assessed cluster quality through silhouette scores and explored within-cluster correlations and distributions.	silhouette_score