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