Programming for Data Science Programming Lab Experimet-7

Course: BCSE207P – Programming for Data Science
Total Marks: 100
Time Allotted: 100 minutes

The Urban Pulse Project

You are a data analyst at "The Urban Pulse Project," a non-profit research initiative studying quality of life in major global cities. Your manager has provided you with a new dataset, `city_metrics.csv`, collected from a variety of public sources. This dataset contains a range of socio-economic and environmental indicators for 50 cities across the world for the year 2023.

Your primary task is to perform an **Exploratory Data Analysis (EDA)** through **data visualization** to uncover patterns, relationships, and stories hidden within this data. The board of directors, who are not data scientists, will review your visualizations. Therefore, the **clarity, aesthetics, and communicative power** of your plots are just as important as the code that generates them.

The Dataset: `city_metrics.csv`

The dataset contains the following variables:

- `city`: Name of the city (Character)
- `country`: Country of the city (Character)
- `continent`: Continent (Africa, Asia, Europe, North America, South America, Oceania) (Factor)
- `population_millions`: Metropolitan population, in millions (Numeric)
- `density_km2`: Population density per square kilometer (Numeric)
- `median_age`: Median age of the population (Numeric)
- `gdp_per_capita_usd`: Gross Domestic Product per capita in US Dollars (Numeric)
- `public_transit_score`: A score from 1-100 rating the quality and availability of public transportation (Numeric)
- `green_space_pct`: Percentage of the city area dedicated to parks and natural reserves (Numeric)
- `air_quality_index`: Annual average Air Quality Index (Higher is worse) (Numeric)
- `avg_commute_time_min`: Average one-way commute time to work in minutes (Numeric)

- `happiness_index`: A composite score of citizen well-being (0-10, higher is better) (Numeric)

Lab Tasks and Questions

Section A: Data Preparation & Initial Exploration (15 Marks)

1. Load the dataset into a dataframe named `city_data`. Check its structure and summary statistics. **(5 Marks)**
2. Are there any missing (`NA`) values in the dataset? If yes, justify your strategy for handling them and write the code to execute it. **(5 Marks)**
3. Create a new factor variable called `size_category` based on `population_millions`:
   - `Small`: < 5 million
   - `Medium`: 5 - 10 million
   - `Large`: > 10 million

   Add this new column to the `city_data` dataframe. **(5 Marks)**

Section B: Univariate & Bivariate Visualizations (40 Marks)

For each plot, ensure you provide appropriate titles, axis labels, and legends. Customize the aesthetics (e.g., colors, themes) for clarity and visual appeal.

4. **Distribution Plot:** Visualize the distribution of the `gdp_per_capita_usd` variable. Use a histogram and a density plot. Based on the plots, comment on the skewness and modality of the distribution. **(10 Marks)**
5. **Categorical Comparison:** Compare the average `happiness_index` across different `continent`s. Choose a single, appropriate plot type (e.g., bar plot, box plot) that effectively shows the central tendency and spread of the data for each category. Justify your choice of plot. What preliminary observation can you make? **(15 Marks)**
6. **Relationship Plot:** Investigate the relationship between a city's wealth (`gdp_per_capita_usd`) and the environmental metric `air_quality_index`. Create a scatter plot. Describe the nature (form, direction, strength) of the relationship you observe. **(15 Marks)**

Section C: Multivariate & Advanced Visualizations (35 Marks)

7. **Multivariate Analysis:** Expand on the scatter plot from question 6. Find a way to incorporate a **third variable** into the same plot. You must choose one of the following to incorporate:

- `size_category` (using `color` or `shape` aesthetics)
- `public_transit_score` (using `size` aesthetic)
- `continent` (using `facets`)

Explain what this new visualization reveals that the previous bivariate plot did not. **(20 Marks)**

8. **Interactive Visualization:** Create an **interactive scatter plot** (using `plotly` or `ggplotly`) using the `happiness_index` on the y-axis and `green_space_pct` on the x-axis. The tooltip should display the `city`, `country`, and both index values when hovered over. Briefly state one advantage of an interactive plot in this context. **(15 Marks)**

Section D: Critical Analysis & Reporting (10 Marks)

9. Based on all your visualizations, write a short summary (max 150 words) directed at the board of directors. Identify **two** of the most interesting or surprising insights you discovered about what factors might be associated with a higher quality of life (as measured by the `happiness_index`) in global cities. Support your claims with references to your visualizations. **(10 Marks)**

Submission Instructions

Rubrics

Section	Task	Excellent (100%)	Proficient (80%)	Developing (60%)	Poor (0-40%)
A (15)	Data Loading & Wrangling	Code is efficient, correct, and handles NA values	Code is mostly correct but may have minor inefficienci	Code has errors but shows some understanding of the	Code is incorrect or missing.

Section	Task	Excellent (100%)	Proficient (80%)	Developing (60%)	Poor (0-40%)
		appropriately. New factor is created correctly.	es or incomplete handling.	required functions.	
B (40)	Plot Creation & Aesthetics	Plots are correct, highly informative, and use appropriate labels, titles, and appealing themes.	Plots are correct but may lack polish (e.g., default titles, unappealing colors).	Plots are generated but contain errors in mapping or lack essential labels.	Plots are incorrect or not generated.
B (40)	Interpretation & Analysis	Observations are accurate, insightful, and use correct statistical terminology.	Observations are correct but may be superficial or lack depth.	Observations are incomplete or contain minor inaccuracies.	Observations are missing or fundamentally flawed.
C (35)	Advanced Visualizations	Multivariate plot is cleverly chosen and	Multivariate plot is adequate but could	Plots are attempted but contain errors in	Plots are incorrect or not attempted.

Section	**Task**	**Excellent (100%)**	**Proficient (80%)**	**Developing (60%)**	**Poor (0-40%)**
		effectively reveals deeper patterns. Interactive plot is functional and well-formatted.	be more effective. Interactive plot works but is basic.	mapping or code execution.	
C (35)	Justification	Justification for the chosen multivariate technique is clear and convincing.	Justification is provided but is weak or unclear.	Justification is attempted but incorrect.	No justification is provided.
D (10)	Summary & Insight	Summary is concise, professional, and highlights two strong, well-supported insights.			