**Homework Assignment: Data Visualization and Manipulation with Python**

**Course:** COMP4010/5120 Data Visualization  
**Deadline:** [Set a due date]  
**Submission Type:** Individual Report (Jupyter Notebook + PDF)

**Overview**

This assignment focuses on **data manipulation and visualization using Python**. You will work with an **IKEA furniture dataset**, exploring various techniques to clean, transform, and visually represent the data.

You will complete **seven tasks**, each requiring you to write **Python code** to perform **data processing and visualization**.

Ensure your code is **efficient, well-commented, and follows best practices**.

**🔹 Dataset: IKEA Furniture Dataset**

The dataset, **ikea\_data.csv**, contains information about various IKEA furniture items, including their names, categories, prices, dimensions, and designers.

**📌 Data Dictionary**

| **Variable** | **Type** | **Description** |
| --- | --- | --- |
| item\_id | Integer | Unique item ID *(not needed for analysis)* |
| name | String | Commercial name of the item |
| category | String | The furniture category |
| price | Float | Current price in **Saudi Riyals (SAR)** |
| old\_price | String | Price before discount in **SAR** |
| sellable\_online | Boolean | Whether the item is sellable online |
| link | String | Link to the product |
| other\_colors | String | Whether other colors are available (Yes/No) |
| short\_description | String | Short description of the item |
| designer | String | Designer(s) of the item |
| depth | Float | Depth of the item (cm) |
| height | Float | Height of the item (cm) |
| width | Float | Width of the item (cm) |

**📌 Submission Requirements**

You should submit your work in **two formats**:

1. **Jupyter Notebook (.ipynb)** – containing your Python code and explanations.
2. **PDF Report** – exported from your notebook, showing **both code and outputs**.

**⚠️ Important Notes**

* **Non-coding questions** should be answered in markdown cells within the notebook.
* Use **Python libraries like pandas, matplotlib, seaborn, plotly** for data manipulation and visualization.

**📌 Tasks**

**Task 1: Convert Prices to USD**

Convert the **price column** from **Saudi Riyals (SAR) to USD**, using an exchange rate of **1 SAR = 0.27 USD**.  
✅ **Create a new column** price\_usd that contains the prices in **USD**.

**Task 2: Split Multiple Designers into Separate Rows**

The **designer column** may contain **multiple designers separated by /**.  
✅ **Transform the dataset** so that each designer appears in a separate row with the same item details.

**Example Input:**

| **item\_id** | **designer** | **price** |
| --- | --- | --- |
| 1 | Designer A | 100 |
| 2 | Designer B/Designer C/Designer D | 200 |

**Expected Output:**

| **item\_id** | **designer** | **price** |
| --- | --- | --- |
| 1 | Designer A | 100 |
| 2 | Designer B | 200 |
| 2 | Designer C | 200 |
| 2 | Designer D | 200 |

**Task 3: Get Top 20 Designers by Number of Items**

Based on the dataset from **Task 2**, find the **top 20 designers** with the most products.  
✅ **Exclude "IKEA of Sweden"** and **remove missing values** from the designer column.

✅ Create a **dataframe** with two columns:

1. designer – Name of the designer.
2. num\_items – Number of items designed.

📌 **Questions to Answer in Markdown:**

1. Who are the **top 3 designers** based on the number of items?
2. What is the number of items designed by **IKEA of Sweden** or by **unknown designers**?
3. Why should we exclude **IKEA of Sweden** from this analysis?

**Task 4: Price Distribution per Designer (Box Plot)**

✅ Create a **box plot** showing the **distribution of item prices (USD)** for the **top designers** identified in Task 3.

* **X-axis:** Designers
* **Y-axis:** Prices in USD

📌 **Questions to Answer in Markdown:**

1. **Describe key findings** from this plot. What do you notice about price distributions?
2. In your opinion, is this an effective visualization? **How could it be improved?**

**Task 5: Distribution of Items per Category (Lollipop Chart)**

✅ Count the number of items in each **furniture category** and visualize the distribution using a **lollipop chart**.

* **X-axis:** Categories
* **Y-axis:** Number of items

📌 **Questions to Answer in Markdown:**

1. What are the **most common categories**?
2. How could this visualization be improved?

**Task 6: Average Price per Category (Lollipop Chart)**

✅ Calculate the **average price (USD) per furniture category** and visualize the distribution using a **lollipop chart**.

* **X-axis:** Categories
* **Y-axis:** Average price in USD

📌 **Questions to Answer in Markdown:**

1. What are the **most and least expensive categories**?
2. Are there any surprising findings?

**Task 7: Price vs. Volume Relationship (Scatter Plot)**

✅ Calculate the **volume of each item** using:

Volume=Height×Width×Depth\text{Volume} = \text{Height} \times \text{Width} \times \text{Depth}Volume=Height×Width×Depth

✅ Create a **scatter plot** showing the relationship between **price (USD)** and **volume**.

* **X-axis:** Price (USD)
* **Y-axis:** Volume
* **Color:** Category

📌 **Questions to Answer in Markdown:**

1. Are **larger items always more expensive**? Why or why not?
2. What trends or outliers do you notice?
3. How could this visualization be improved?

**Task 8: Price Prediction Using Regression Models (For PhD students taking COMP5120)**

In this task, you will apply **regression models** to predict **product prices** based on the dataset.

✅ **Steps:**

1. **Feature Selection**
   * Select relevant features that could predict the price (e.g., **category, volume, designer, sellable\_online**).
   * Encode categorical variables using **one-hot encoding**.
2. **Train a Regression Model**
   * Split the data into **training (80%) and testing (20%) sets**.
   * Train a **Linear Regression model** to predict the price in USD.
   * Evaluate the model's **performance (R² score, MAE, RMSE)**.
3. **Experiment with Other Models**
   * Try at least **one alternative model** (e.g., **Random Forest Regression, Gradient Boosting, or Ridge Regression**) and compare results.
   * Which model performs best, and why?

📌 **Questions to Answer in Markdown:**

1. Which features were **most important** for price prediction?
2. How well does **Linear Regression** perform? What are its limitations?
3. Which **alternative model** performed better, and why?
4. What additional **features** could improve the model’s accuracy?

**📌 Bonus Task (Optional)**

✅ Use **interactive visualizations (e.g., Plotly, Altair)** to make the plots **more dynamic**.  
✅ Try **clustering designers by pricing patterns** using **K-Means clustering**.

**📌 Submission Checklist**

🔲 **Jupyter Notebook (.ipynb)** with all code, explanations, and visualizations.  
🔲 **PDF Export** of your notebook, including code outputs and written answers.  
🔲 **Ensure plots are clear, properly labeled, and easy to interpret.**  
🔲 **Write a short summary (1-2 paragraphs) about the insights you gained from this analysis.**

**📌 Grading Criteria**

✔ **Preprocessing & Cleaning (20%)** – Correct handling of missing data, splitting designers, and currency conversion.  
✔ **Data Transformation (20%)** – Accurate calculations for category counts, averages, and volume.  
✔ **Visualization & Interpretation (30%)** – Effective use of **box plots, lollipop charts, and scatter plots**.  
✔ **Clarity & Explanation (30%)** – Insightful observations, well-documented code, and strong report structure.

**📢 Final Reminder:**

This homework is **individual work**. **Plagiarism is strictly prohibited**—all work must be your own.

🚀 **Good luck!** 🚀