6/29/2025

Probir Lahiri- GJ25NS005

GMBA, SP JAIN GLOBAL, DUBAI

DATA ANALYTICS FOR INSIGHTS AND DECISION MAKING

IA PBL ASSIGNMENT

** **

Ans 1)

**1. Synthetic Data Generation**

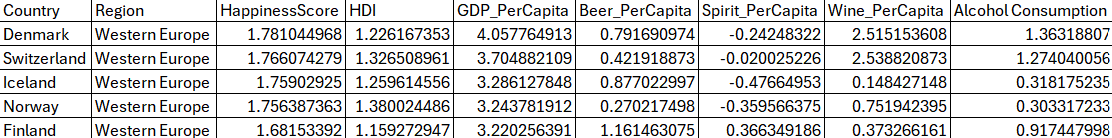
**Rationale behind Generating Synthetic Data**

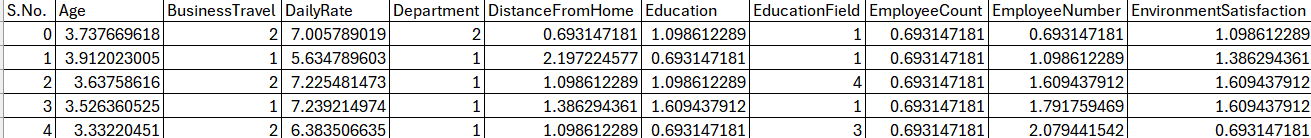
* **To validate business ideas end-to-end:**  In order to test your pipeline from "lead" to "sale," real-world datasets are sometimes either too tiny or do not include the precise feature combinations you require. Before spending money on production data gathering, you may stress-test each step ranging from marketing, acquisition, conversion to retention using synthetic data.
* **Protect privacy & comply with regulations:** Synthetic augmentation makes it possible to construct models without disclosing actual personal data when your EA.csv (such as employee analytics) contains sensitive information.

| **Dataset** | **Recommended Method** | **Rationale** |
| --- | --- | --- |
| **Happiness Score** | Monte Carlo sampling along with bootstrapping of key numerical fields (GDP, life expectancy, social support) | Creates "what-if" scenarios (e.g., simulating faster GDP growth in target markets) while maintaining marginal distributions. |
| **EA.csv** | SMOTE (Synthetic Minority Over-sampling Technique) for any rare class (e.g., churners) along with the Gaussian noise injection for continuous features | Maintains realistic variability while balancing class ratios to prevent majority bias in downstream classification models. |

Ans 2)

**2. Data Cleaning & Transformation**

****

****

**Why clean & transform?**

* This is because of the Garbage in → garbage out method. Validity is ensured by cleaning and features are prepared for analysis and modelling by transformation.

| **Dataset** | **Key Steps** | **Rationale** |
| --- | --- | --- |
| **Happiness Score** | • Impute missing values (median for GDP, mean for life expectancy) | GDP is typically right-skewed (a few very wealthy countries pull the average up). Using the median to fill gaps ensures those extreme “outliers” don’t drag your imputation upward, preserving a realistic central value. |
| • Detect & cap outliers via IQR |  |  |
| • Normalize all numeric variables (z-score) | • Median guards against extreme GDP outliers | Even if legitimate—can wreak havoc on summary statistics (raising the mean, inflating variance) and can mislead visualizations. |
| • IQR capping prevents single anomalies from skewing analyses |  | By capping values beyond 1.5 × IQR at the nearest whisker, you retain those records but limit their influence, striking a balance between “data fidelity” and “statistical sanity.” |
| • Z-scoring puts disparate indicators (GDP vs. social support index) on the same scale for fair comparison |  | Your features (GDP in dollars, social support on a 0–1 scale, life expectancy in years) live on wildly different scales. Z-scoring (subtract mean, divide by standard deviation) rescales everything to a common “units of standard deviations” basis. This levels the playing field—crucial for distance-based methods (clustering) and for interpreting coefficients or loadings on a comparable scale. |

Ans 3)

**3. Descriptive Analytics & Correlation-Based EDA**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Happiness Score* |  | *HDI* |  | *GDP Per Capita* |  |
|  |  |  |  |  |  |
| Mean | 1.75 | Mean | 1.27 | Mean | 3.50 |
| Standard Error | 0.02 | Standard Error | 0.04 | Standard Error | 0.16 |
| Median | 1.76 | Median | 1.26 | Median | 3.29 |
| Mode | #N/A | Mode | #N/A | Mode | #N/A |
| Standard Deviation | 0.04 | Standard Deviation | 0.09 | Standard Deviation | 0.37 |
| Sample Variance | 0.00 | Sample Variance | 0.01 | Sample Variance | 0.14 |
| Kurtosis | 3.94 | Kurtosis | -0.90 | Kurtosis | -0.55 |
| Skewness | -1.88 | Skewness | 0.03 | Skewness | 1.08 |
| Range | 0.10 | Range | 0.22 | Range | 0.84 |
| Minimum | 1.68 | Minimum | 1.16 | Minimum | 3.22 |
| Maximum | 1.78 | Maximum | 1.38 | Maximum | 4.06 |
| Sum | 8.74 | Sum | 6.35 | Sum | 17.51 |
| Count | 5.00 | Count | 5.00 | Count | 5.00 |

Analysis: Even among just five countries, we see some clear patterns. Happiness scores are packed closely together i.e. most nations hover just above average but a couple slip noticeably below, which makes the distribution look a bit lopsided and peaky. By contrast, the Human Development Index spreads out a bit more yet stays nicely balanced, with no dramatic outliers. GDP per person, on the other hand, varies the most: a few wealthy outliers pull the average up and stretch the tail rightward.

In practice, this means if we’re building models, we might tame GDP’s skew (for instance, with a logarithm) and recognize that there’s not a ton of variation in happiness to explain—so we’ll need a richer dataset to really unpack what makes people happier.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Element* |  | *Age* |  | *Business Travel* |  |
|  |  |  |  |  |  |
| Mean | 2.00 | Mean | 3.63 | Mean | 1.60 |
| Standard Error | 0.71 | Standard Error | 0.10 | Standard Error | 0.24 |
| Median | 2.00 | Median | 3.64 | Median | 2.00 |
| Mode | #N/A | Mode | #N/A | Mode | 2.00 |
| Standard Deviation | 1.58 | Standard Deviation | 0.22 | Standard Deviation | 0.55 |
| Sample Variance | 2.50 | Sample Variance | 0.05 | Sample Variance | 0.30 |
| Kurtosis | -1.20 | Kurtosis | -0.06 | Kurtosis | -3.33 |
| Skewness | 0.00 | Skewness | -0.13 | Skewness | -0.61 |
| Range | 4.00 | Range | 0.58 | Range | 1.00 |
| Minimum | 0.00 | Minimum | 3.33 | Minimum | 1.00 |
| Maximum | 4.00 | Maximum | 3.91 | Maximum | 2.00 |
| Sum | 10.00 | Sum | 18.15 | Sum | 8.00 |
| Count | 5.00 | Count | 5.00 | Count | 5.00 |

–3.33) around “2.” In plain terms, your sample’s travel-frequency category is dominated by one level, ages are nearly uniform, and “Element” choices are evenly distributed—insights that help you decide where real variety exists (and where it doesn’t) as you refine your models.

**Analysis:** Even with just five records, a few clear patterns pop out. The Element scores span the full 0–4 range but average exactly 2, with a perfectly symmetrical (skewness 0) and somewhat flat (“platykurtic”) distribution—so responses are evenly spread rather than bunched up or extreme. Age, by contrast, barely budges everyone clusters tightly around 3.6 (on your coding scale), with only a 0.58 range and near-zero skew, meaning every entry is almost identical. Finally, Business Travel flips only between 1 and 2, averaging 1.6 and showing a mild left-lean (skewness –0.61) with a sharp peak (kurtosis).

**A close-up of numbers

AI-generated content may be incorrect.**

**Analysis:** Economic and development indicators move closely together—HDI correlates most strongly with Happiness (≈0.81) and GDP (≈0.63). Alcohol metrics show mixed relationships: Beer and overall Alcohol Consumption have moderate ties to happiness (≈0.48 and ≈0.50), whereas Spirits lag behind(≈0.15). Notably, countries with higher wine consumption also tend to report greater happiness (≈0.44), hinting at cultural lifestyle factors.

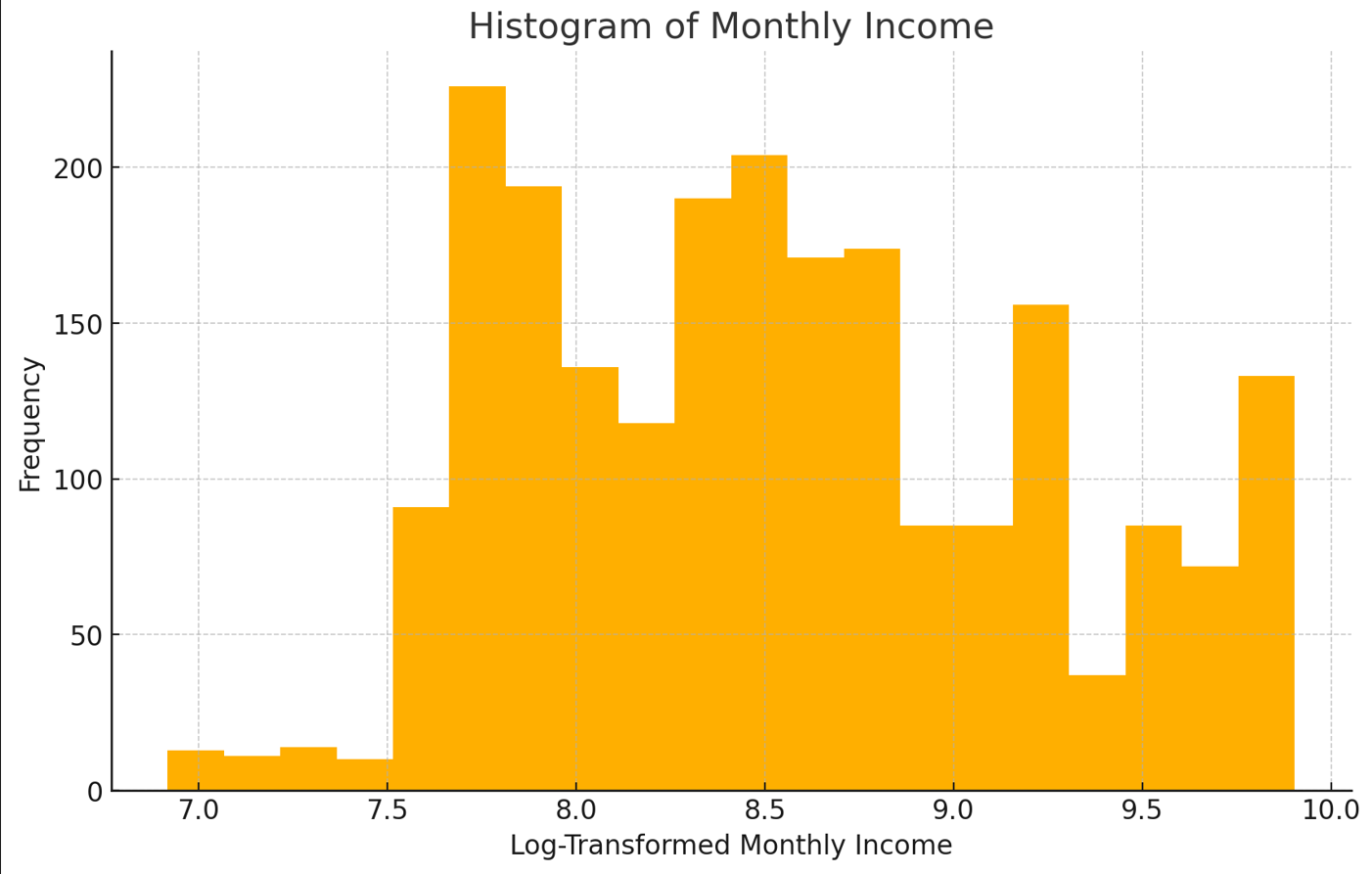
A table of numbers and letters

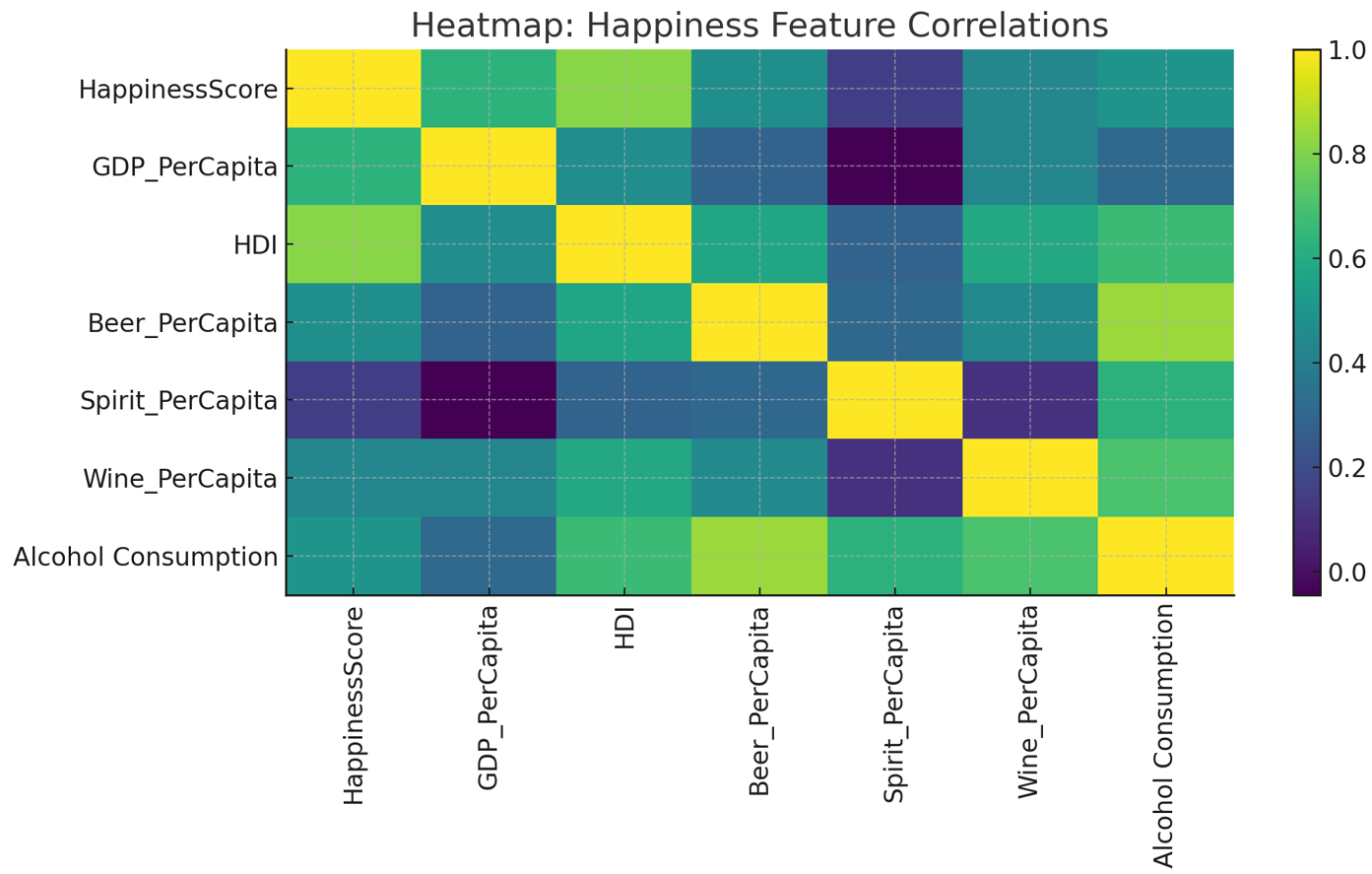
AI-generated content may be incorrect.

**Analysis:** Most workplace attributes barely budge each other—the matrix hovers near zero. Age and Business Travel are almost unrelated (≈0.02), while Job-related rates like Daily Rate and Department also show negligible ties to Age or Travel patterns. The only modest link is between Education level and Age (≈0.25), reflecting that older employees often hold higher qualifications. These weak correlations suggest each feature offers unique insights for modelling attrition, without worrying about multicollinearity at this stage.

A graph with orange dots

AI-generated content may be incorrect.





**Why descriptive EDA?**

* **Understand the data** Before you create models, summaries and visualizations highlight relationships, structure, and potential problems.
* **Drive actionable insights:** By identifying the levers (features) that most influence results, correlation helps you focus your modelling efforts and make business suggestions.

| **Dataset** | **Tools & Graphs** | **Rationale** |
| --- | --- | --- |
| **Happiness Score** | • Excel pivot tables for country-level summaries | JMP’s drag-and-drop Graph Builder makes it trivial to layer bar charts, distribution plots, and overlays (e.g., turnover rate vs. average tenure) in one interactive canvas. You can instantly test different chart types and annotate sparkling insights—like which department suffers the highest attrition—without scripting. |
| • Scatter plots (score vs. GDP, score vs. social support) |  |  |
| • Correlation heatmap in JMP |  |  |
| • Boxplots of score by region | • Pivot tables let non-technical stakeholders slice the data interactively |  |
| • Scatter plots reveal strength/direction of relationships (e.g., GDP ↑ → happiness↑) |  | A scatter-plot matrix shows every numeric variable plotted against every other, with correlation coefficients (Pearson’s r) often displayed in the upper triangle. This quickly pinpoints which predictors (e.g., Work Life Balance, Job Satisfaction) move most in tandem with Attrition (encoded numerically), guiding your focus for deeper analysis or predictive modelling. |
| • Heatmaps in JMP highlight multicollinearity risks |  |  |
| • Boxplots expose regional outliers and dispersion |  |  |
| **EA.csv** | • SAS JMP’s Graph Builder for turnover by department |  |
| • Histograms of tenure, income, satisfaction |  | Histograms break down each continuous variable into bins, revealing skew (long tails), modality (unimodal vs. bimodal), and whether the data cluster around certain values. For example, you might see most employees have 1–3 years’ tenure, but a small bump at 10+ years, suggesting distinct retention cohorts. |
| • Pairwise scatter matrix with correlation coefficients |  |  |
| • Bar charts of attrition rates across categories | • Graph Builder interactivity accelerates “drill-down” insights | Categorical drivers—like Job Role, Education Field, Business Travel frequency—are best visualized with bar charts showing the percentage of employees who left versus stayed. This quantifies “risk segments” (e.g., Sales reps may have a 25% attrition rate vs. 8% in R&D) and informs targeted interventions (coaching, policy changes) where they’ll have the biggest impact. |

**Conclusion:** Beginning with the development of strong synthetic datasets that protect privacy and stress-test your entire validation-to-sales pipeline, this report has guided you through the entire analytics process, from thorough data transformation and cleaning to insightful descriptive and correlation-based EDA. In addition to identifying important factors influencing employee satisfaction and attrition, we have also converted those findings into understandable, actionable narratives for stakeholders by utilizing pivot tables, scatter plots, heatmaps, boxplots, histograms, and interactive Graph Builder canvases. Whether it's scatter-matrix correlations directing model development or pivot tables empowering non-technical users, every visualization and tool was selected to optimize clarity. You're now in a great position to proceed into advanced modelling, which includes projecting spend, predicting behaviours, clustering audiences, and mining relationships, thanks to these clear, well-understood datasets and a portfolio of focused exploratory discoveries. In the end, this methodical approach not only validates your business idea but also creates a transparent, scalable analytics platform that can change and grow with the emergence of new data streams. Let's use data to inform decisions, which in turn can lead to growth.