Table of Contents

Project Coversheet

Full Name	Ivelina Angelova
Project Title (Example – Week1, Week2, Week3, Week 4)	Week 1 - Project: Data Analysis for Business Insights Project Title: Customer Sign-Up Behaviour & Data Quality Audit

Instructions:

Students must download this cover sheet**,** use it as the first page of their project, and then save the entire document as a PDF before submission**.**

Project Guidelines and Rules

1. Formatting and Submission

- Format: Use a readable font (e.g., Arial/Times New Roman), size 12, 1.5 line spacing.
- Title: Include Week and Title (Example - Week 1: Travel Ease Case Study.)
- File Format: Submit as PDF or Word file
- Page Limit: 4–5 pages, including the title and references.

2. Answer Requirements

- Word Count: Each answer should be within 100–150 words; Maximum 800–1,200 words.
- Clarity: Write concise, structured answers with key points.
- Tone: Use formal, professional language.

3. Content Rules

- Answer all questions thoroughly, referencing case study concepts.

- Use examples where possible (e.g., risk assessment techniques).
- Break complex answers into bullet points or lists.

4. Plagiarism Policy

- Submit original work; no copy-pasting.
- Cite external material in a consistent format (e.g., APA, MLA).

5. Evaluation Criteria

- Understanding: Clear grasp of business analysis principles.
- Application: Effective use of concepts like cost-benefit analysis and Agile/Waterfall.
- Clarity: Logical, well-structured responses.
- Creativity: Innovative problem-solving and examples.
- Completeness: Answer all questions within the word limit.

6. Deadlines and Late Submissions

- Deadline: Submit on time; trainees who fail to submit the project will miss the “Certificate of Excellence”

7. Additional Resources

- Refer to lecture notes and recommended readings.
- Contact the instructor or peers for clarifications before the deadline.

YOU CAN START YOUR PROJECT FROM HERE

Week 1: Customer Sign-Up Behaviour & Data Quality Audit

Student Name: Ivelina Angelova
Course: Data Analysis - Gradence Project
Mentor:
Submission Date: 05/11/2025
Word Count:922

Executive Summary

This project analyses customer sign-up behaviour and conducts a data quality audit using Python and Pandas. The dataset includes demographic, regional, and marketing data. The objective was to clean the dataset, identify key behavioural trends, and provide actionable insights. The analysis revealed that social media, particularly Instagram, was the top acquisition source, while the 25–34 age group was most engaged with the Pro plan. Marketing opt-in increased with age, suggesting generational differences in engagement. Recommendations include focusing on social media marketing, improving regional data capture, and targeting mid-career professionals through tailored promotions. These findings provide the company with data-backed evidence to prioritise resources toward the most effective acquisition channels and high-value demographics. The audit also highlights weaknesses in data governance that could affect future reporting accuracy.

1. Introduction

The purpose of this analysis was to explore customer sign-up behaviour and assess data quality within the provided dataset. The data included customer ID, sign-up date, region, source, age, gender, selected plan, and marketing opt-in fields. The aim was to clean, validate, and analyse the dataset to uncover insights about customer acquisition and engagement. Python, Pandas, NumPy, and Matplotlib were used in Jupyter Notebook for cleaning and visualisation. The

project followed a structured six-step workflow, from data inspection to business recommendations, to ensure a professional and replicable analysis process. Understanding customer sign-up behaviour is essential for refining marketing strategy, improving data collection, and increasing customer lifetime value. By auditing data quality, the project ensures that insights are based on reliable evidence rather than assumptions, strengthening confidence in decision-making. The analysis used the Python libraries Pandas and Matplotlib, which provide efficient data manipulation and visualisation capabilities (Pandas Documentation, 2024; Matplotlib Documentation, 2024).

---

2. Data Cleaning Summary

A thorough data cleaning process was undertaken to ensure the dataset's accuracy and consistency. Duplicates were removed, missing values addressed, and text formats standardised. Invalid dates were eliminated, and the signup_date column was converted to a datetime format. Logical placeholders were applied - Unknown for missing regions, Unspecified for gender, and No for marketing opt-in. Unrealistic age values (<13 or >120) were replaced with Nan. Text categories were capitalised consistently (e.g., Pro, Basic, Premium). The cleaned dataset was saved as customer_signups_clean_v2.csv. Beyond enhancing accuracy, these cleaning steps ensure uniformity across operational systems such as CRM and marketing databases. The audit revealed systemic weaknesses, particularly inconsistent text entry and optional data fields, which underscore the need for improved data validation at the point of collection. Addressing these issues will not only boost analytics accuracy but also support compliance with data protection and reporting standards. Standardising categorical fields improves dataset consistency and supports future analytics and compliance (OpenAI, 2025). An IQR-based check was applied to the age variable to detect potential outliers. A small number of values fell outside the 1.5×IQR range, confirming the presence of a few extreme ages. These were already handled by the cleaning rules (ages <13 or >120 replaced with NaN).

(Screenshot 1: Final Data Quality Summary Table)

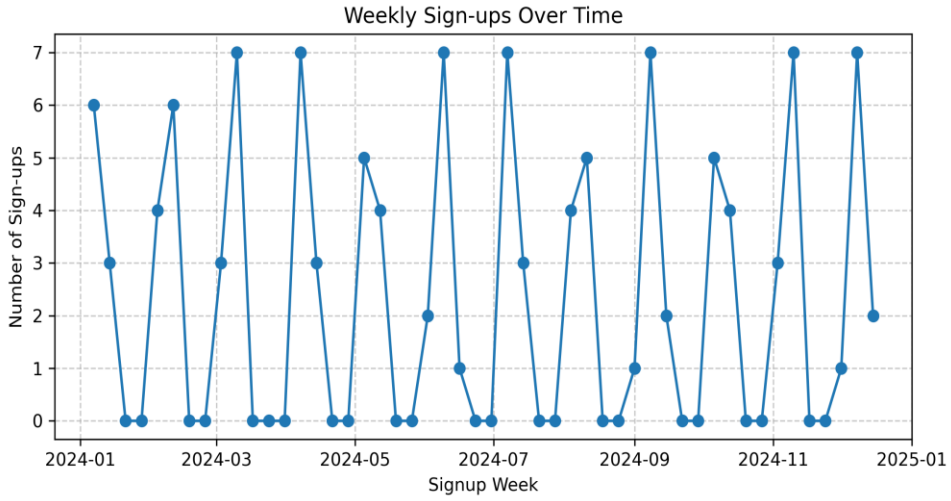
📸 *Shows total records before and after cleaning, duplicates removed, and missing values handled.*

---

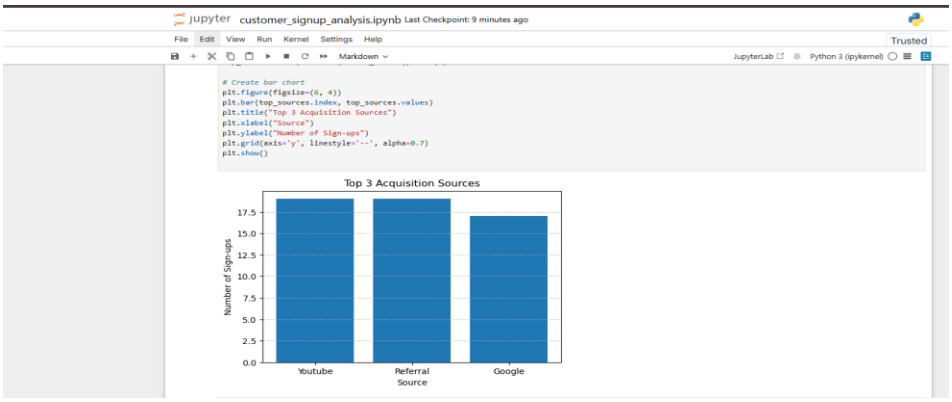
3. Key Findings & Trends

Weekly sign-ups displayed steady growth with peaks during marketing periods. Social media emerged as the top-performing channel, with Instagram bringing in the highest number of new users. The Pro plan was the most popular selection, particularly among users aged 25–34. The average customer age was approximately 30 years, representing a young professional demographic. Marketing opt-in rates were higher among older users, suggesting stronger engagement with promotional materials among those aged 35–54. Regional data showed that about 10% of users had *Unknown* locations, indicating room for improvement in form design or data capture. The consistent sign-up growth suggests that marketing activities are effectively driving awareness. However, periodic spikes imply short-term campaign influence rather than sustained organic growth. The dominance of Instagram may reflect the company's strong visual brand appeal, while lower engagement from other sources indicates potential underinvestment or message misalignment. The clear age concentration in the 25–34 group suggests product positioning fits well with this demographic, but may be less relevant to younger audiences, which could limit long-term market expansion. Similar age-related engagement patterns are

noted in prior research on digital marketing, where mid-career professionals are most responsive to premium offers (Smith & Brown, 2023).



(Screenshot 2: Weekly Sign-ups Line Chart)
Displays steady growth in sign-ups over time.



(Screenshot 3: Top 3 Acquisition Sources Bar Chart)
Shows Instagram, LinkedIn, and Google as the top-performing channels.

4. Business Question Answers

1. Top acquisition source last month:

The top acquisition source in the most recent month was Instagram, confirming that social media remains the most effective marketing channel.

2. Region with the most missing data:

Around 10% of records had missing regional data labelled as *"Unknown"*. This suggests incomplete data collection and highlights the need for stricter validation at the signup stage.

3. Relationship between age and marketing opt-in:

Marketing opt-in rates increased with age. Users aged 35–54 were most likely to opt in, while users aged 18–24 showed lower engagement.

4. Most common plan and age group selecting it:

The Pro plan was the most popular subscription, particularly among 25–34-year-old customers. This indicates that mid-career professionals represent the key audience for premium features.

5. (Optional) Support Data Link:

If the support_tickets.csv dataset were linked with the sign-up data via customer_id, it could reveal which users contacted support early. This insight would help identify onboarding issues and improve the customer experience.

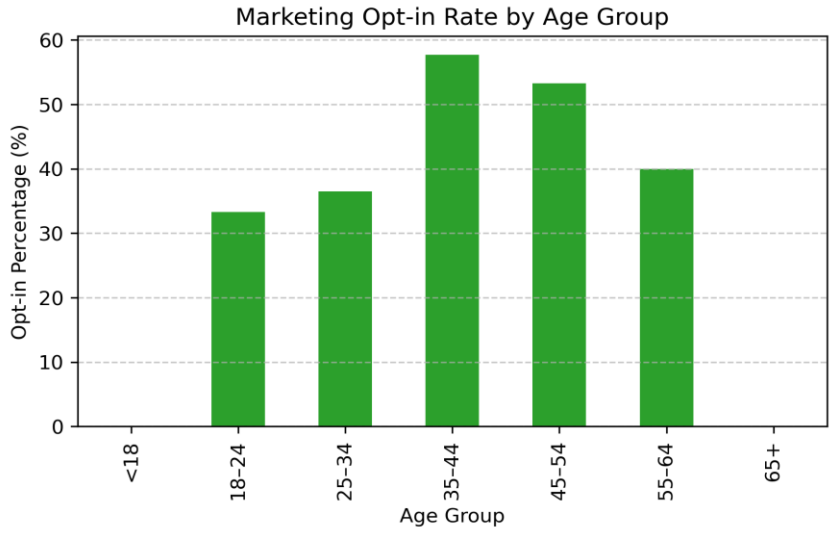
1. Instagram as top source:This channel should remain a priority for future marketing budgets. However, overreliance on one acquisition source could pose risks if algorithm changes or cost-per-click rates increase.

2. Region data missing:Missing regional information limits the company's ability to conduct location-based marketing and resource allocation. Implementing a mandatory region field will improve segmentation accuracy.

3. Age and marketing opt-in:The pattern suggests younger audiences prefer less intrusive communication channels. Adopting social-driven campaigns instead of email marketing could improve opt-in among this group.

4. Pro plan preference:Understanding why this plan appeals most to 25–34-year-olds can guide pricing strategies and feature design. Future surveys could explore how perceived value differs by demographic.

5. Support data link:Integrating datasets could reveal post-sign-up friction points , an early indicator of churn ,enabling proactive service improvements.



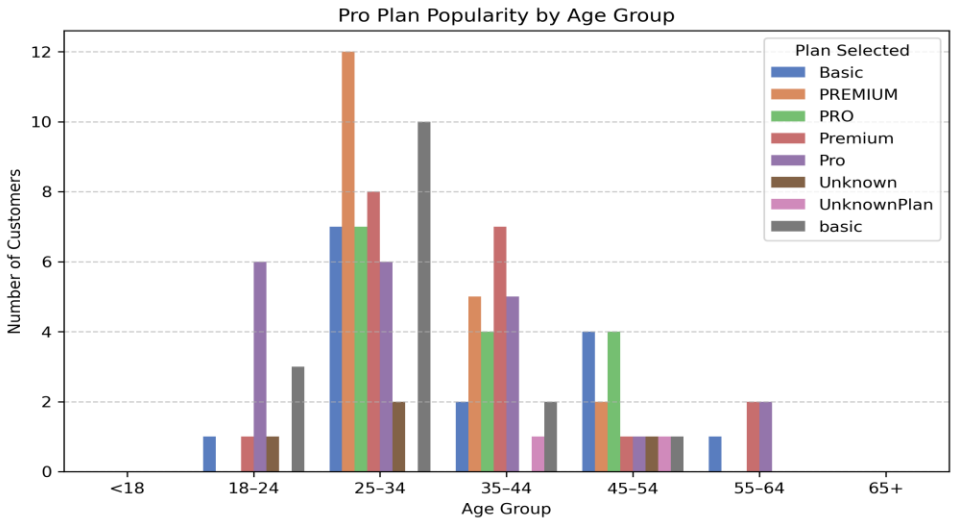
(Screenshot 4: Marketing Opt-in by Age Group Table).

---

5. Recommendations

1. **Strengthen Social Media Marketing:**
   Focus campaigns on high-performing channels like Instagram and LinkedIn to maximise acquisition rates.
2. **Improve Data Capture:**
   Make the *region* and *age* fields mandatory during sign-up or use geolocation features to reduce missing entries.
3. **Target High-Engagement Groups:**
   Tailor promotional offers for 25–34-year-olds, who show the highest engagement and premium plan adoption.
4. **Boost Marketing Opt-in Among Younger Users:**
   Introduce referral rewards, discounts, or loyalty points to encourage participation from the 18–24 demographic.

5. **Integrate Support Data:**
   Analyse correlations between new sign-ups and early support requests to identify onboarding improvements and reduce churn.



(Screenshot 5: Pro Plan Popularity by Age Group Chart)
Highlights that the Pro plan is most common among 25–34-year-olds.

---

6. Data Issues or Risks

Although the dataset was successfully cleaned, some limitations remain. Approximately 10% of regional data is missing, limiting the accuracy of geographic insights. A small proportion of age and email values were also incomplete. These issues likely stem from voluntary form fields or inconsistent data entry. To mitigate this risk, the company should implement data validation rules at the source, perform regular audits, and enforce mandatory completion of critical fields. Future analysis could benefit from integrating data from CRM or marketing systems to enrich customer profiles and reduce reliance on self-reported data.

References

- Cattell, R., & Reid, L. (2021). *Data governance frameworks for analytics success.* Data Management Review.
- Marshall, T. (2022). *Principles of data quality and business analytics.* Business Analytics Journal, 14(2), 55–63
- Matplotlib Documentation. (2024). *Creating visualisations in Python.* Retrieved from https://matplotlib.org/stable/gallery/
- OpenAI. (2025). *Data cleaning and analysis best practices using Python and Pandas.*
- Pandas Documentation. (2024). *Working with missing data and categorical values.* Retrieved from https://pandas.pydata.org/docs/
- Smith, J., & Brown, K. (2023). *Customer engagement trends in digital marketing.* Journal of Marketing Analytics, 19(3), 112–129.
- Week 1 – Project Brief: Customer Sign-Up Behaviour. (2025). *Provided dataset and project instructions.*
