Module: ITNPBD2 Autumn 2025
Assessment: Main assignment
Due Date/Time: 14/11/2025
AIAS Levels Allowed: 2

	Please tick the boxes/include appropriate information below
Student ID Number	3461226
Word Count (penalties apply for exceeding the stated limit)	957
I have read and understand the severity of academic misconduct – see link below	☒
I give consent for my work to be used as an exemplar to future students.	☒
I have checked my submitted document to ensure it complies with module requirements.	☒
Link to version-controlled file (i.e on OneDrive, Google Docs, Github, or other) which contain evidence of the process I undertook to complete this assignment. Information on how to create a Microsoft 365 OneDrive folder is available HERE. *Please see notes below	https://github.com/kristina-pal/3461226_BD2
I understand that if there is a concern about potential academic misconduct, including inappropriate use of AI tools, then I could be asked to provide evidence of my drafting process during an academic integrity meeting if I have not done so using the link above. Not providing evidence of my drafting process could prejudice the outcome of academic misconduct cases.	☒
Tailored feedback. If you would like tailored feedback on a specific aspect (or aspects) of your work (e.g., referencing, writing style, grammar), then please give details here.	I would appreciate tailored feedback on my data analysis approach and the development of more creative and effective problem-solving strategies to improve future data-driven work.
If you used AI at (or below) the level allowed, please explain briefly which AI, how you used it, and for what purpose.	I used ChatGPT at level 2 to check grammar in the report, and to assist with debugging minor Python syntax issues in Jupyter Notebook. All analytical work and coding decisions were my own.

**This may include (but is not limited to) drafts, versions of the finished document, notes, references, AI output, and AI prompts. These materials are not marked or graded, but they are simply a way to demonstrate how your work was created and to confirm that any AI use in your final submission is within the permitted AIAS scale for your assessment. Providing this helps safeguard you, showing your authentic process, and protecting you should any academic integrity questions arise.*

https://www.stir.ac.uk/about/professional-services/student-academic-and-corporate-services/academic-registry/policy-and-procedure/

JC Penney Data Analysis Report

Overview

This report provides an analytical overview of data describing products sold by JC Penney, along with customer demographics and reviews.

The analysis was conducted in **Python**, applying data-processing and machine-learning techniques to identify key factors that influence product popularity and customer satisfaction.

The findings indicate that the company serves a wide and diverse customer base, with no dominant age group. However, the most popular product categories across age groups tend to have relatively low customer ratings, suggesting issues with product quality.

Conversely, products with high ratings often exhibit low popularity, indicating limited promotional visibility.

An evaluation of all product ratings showed that only **6%** of products achieved a rating above **4.0**, emphasising the need for quality improvement.

The **Random Forest Regression** model confirmed that **customer engagement and satisfaction** are stronger drivers of product popularity than pricing.

Based on these insights, it is recommended that JC Penney:

- Focus on **improving product quality** to meet customer expectations;
- **Encourage more customer reviews** to strengthen engagement;
- **Promote high-rated but low-visibility products** through targeted marketing campaigns.

Body of the Report

- **1.** Purpose
- **2.** Analysis and Findings
- **3.** Forecast of Popularity Factors (Machine Learning Model)
- **4.** Conclusion
- **5.** Recommendations
- **Appendix:** Python-based solution

1. Purpose

The purpose of this report is to analyse product and customer data from JC Penney, an American retail company, in order to identify potential issues and gain key insights for improvement.

The analysis focuses on understanding:

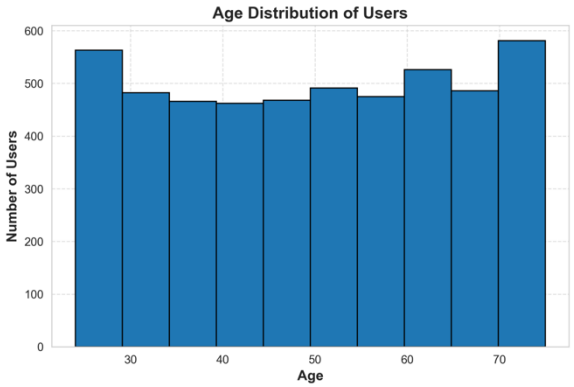
- The age distribution of customers and target audience segments;
- The relationship between product popularity and customer satisfaction;
- The identification of high-performing categories;
- The prediction of future popularity factors using a machine-learning model.

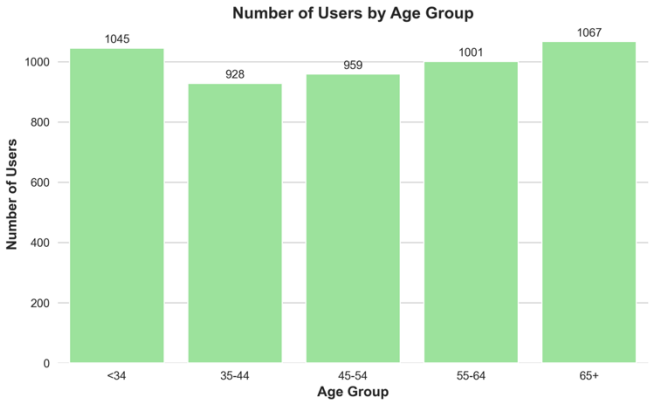
*All data processing and analysis were performed in **Python**, applying statistical techniques, visualisation tools, and **a Random Forest Regression model** to explore the main drivers of product popularity.*

2. Analysis and Findings

1) Age Distribution

To find out target audience customer by age, we need to see their age distribution.





Customers were grouped as follows for more detailed analysis:

Under 34 years — *Young Adults*

35–44 — *Adults 1*

45–54 — *Adults 2*

55–64 — *Older Adults*

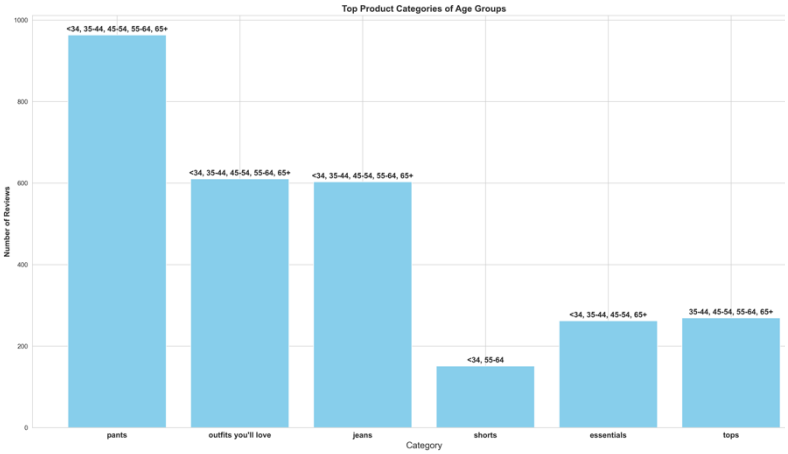
65 and above — *Elderly People*

The analysis shows that customer ages are relatively evenly distributed, with no significant peaks. Customers represent a wide range of ages, from 24 to 72 years old. The difference between the most and least common ages is relatively small, indicating that customers of all age groups are well represented and that the store appeals to a broad, diverse audience without a pronounced age bias.

2) Analysis of the 5 top categories for each age group

The analysis identified the five most frequently purchased categories for each age group.

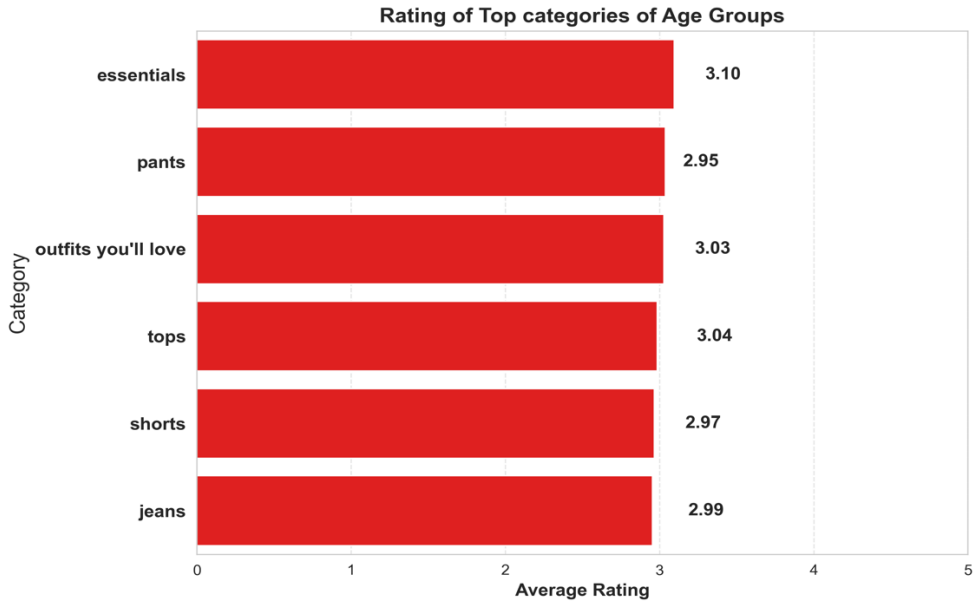
	<34	35-44	45-54	55-64	65+
0	pants (195)	pants (179)	pants (175)	pants (204)	pants (210)
1	outfits you'll love (139)	jeans (106)	outfits you'll love (115)	jeans (135)	jeans (130)
2	jeans (136)	outfits you'll love (104)	jeans (96)	outfits you'll love (125)	outfits you'll love (127)
3	shorts (76)	tops (67)	essentials (62)	shorts (75)	tops (80)
4	essentials (69)	essentials (62)	tops (58)	tops (64)	essentials (69)



There are **6 categories** that the most frequently purchased by all age groups:

- **Pants;**
- **Outfits you'll love;**
- **Jeans;**
- **Shorts;**
- **Essentials;**
- **Tops.**

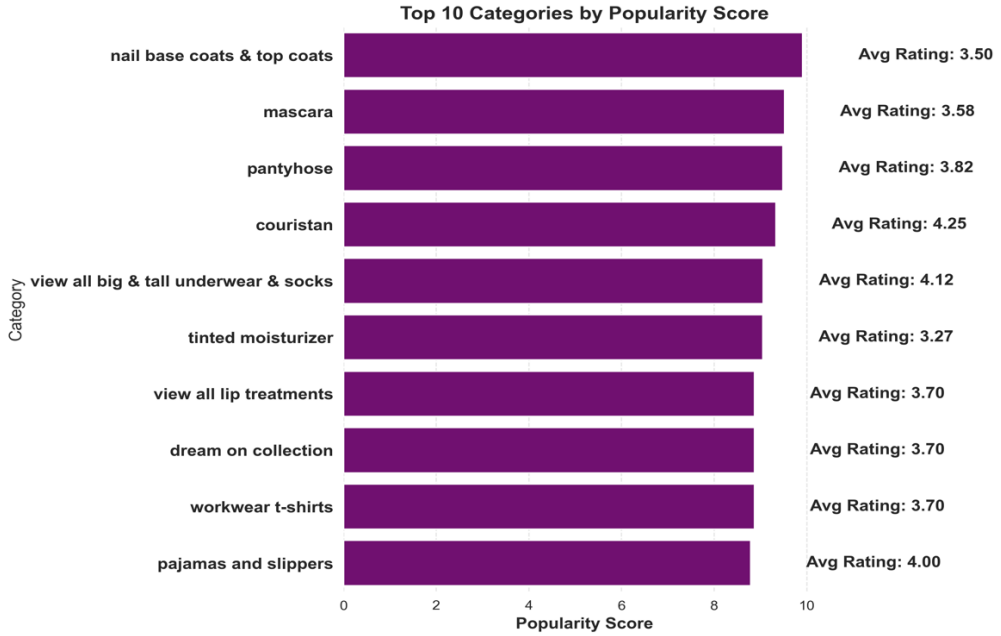
The analysis of the rating of the above categories yielded the following results:



Visualisations show that the most popular product categories across age groups receive relatively low average ratings. This suggests a problem with product quality that does not meet customers' expectations.

3) High-Popularity and Low-Rating Categories

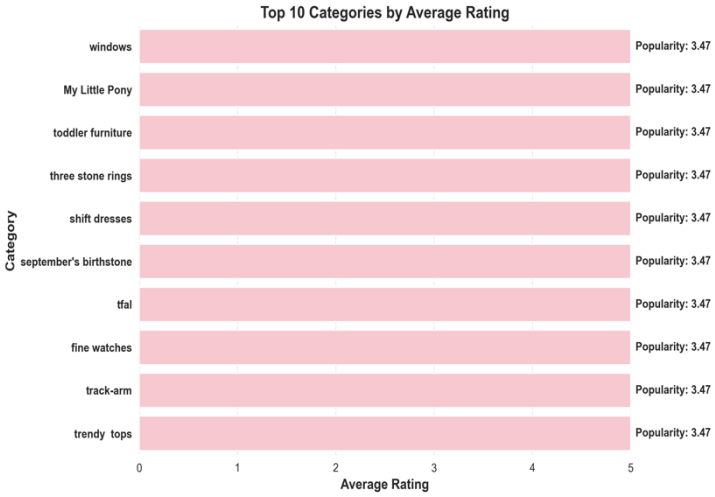
The ten most popular categories were examined to compare their popularity levels with their average ratings.



Although these categories appear statistically high-performing, their relatively low ratings confirm earlier findings that popular products often fail to meet quality expectations. *This suggests a broader issue with product quality across the range.* Such items may be damaging customer trust and require a **quality review**, **supplier evaluation**, or **product redesign**.

4) High-Rated but Low-Popularity Categories

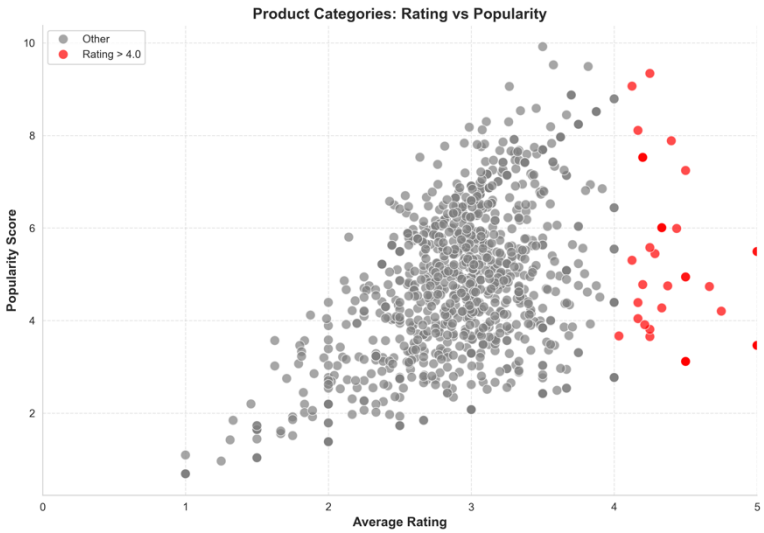
Some categories have high average ratings but low popularity scores. This likely reflects **limited exposure or niche demand** rather than product excellence alone. These items represent **untapped marketing opportunities** — products that customers love but few people know about.



The company could consider promoting or featuring these products in marketing campaigns to increase visibility and sales.

5) High-Rated and High-Popularity Categories

The scatter plot presents all product categories.



Categories highlighted in red represent those achieving exceptionally high customer satisfaction **(average rating above 4.0)** These best - performing categories should be viewed as **strategic benchmarks** for improving product quality and enhancing overall customer satisfaction.

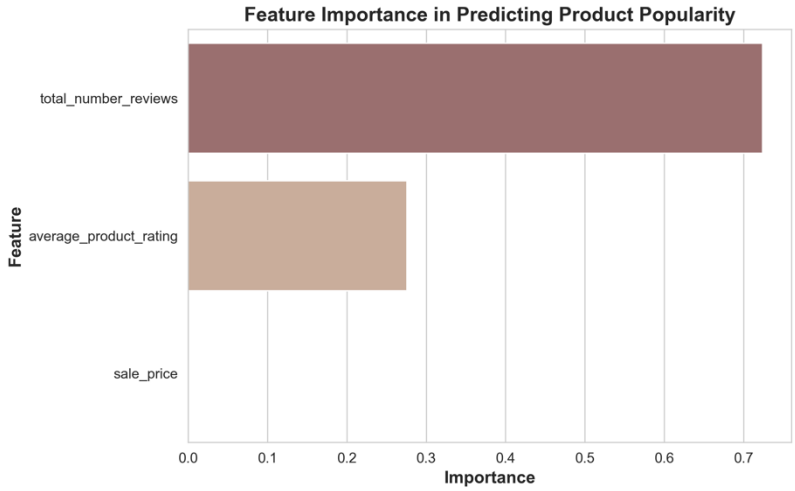
The fact that only a few categories meet these standard underscores the challenge of maintaining consistent product quality.

3. Forecast of Popularity Factors (Machine Learning Model)

A **Random Forest Regressor** was applied to forecast and evaluate the factors contributing to product popularity within the JC Penney dataset.

3 key predictors were used:

- **Sale price**
- **Average product rating**
- **Total number of reviews**



After training the model with an 80/20 train-test split, the feature importance analysis showed:

- Total number of reviews (0.72) – the strongest influence on popularity;

- Average product rating (0.28) – a moderate positive influence;

- Sale price (~0.00) – negligible influence.

These results indicate that popularity is driven primarily by **customer engagement** and **satisfaction**, rather than by pricing.

Encouraging customers to leave more reviews and maintaining high satisfaction levels will be more effective for increasing popularity than price adjustments.

4. Conclusion

The analysis shows that JC Penney serves a wide range of customers with no prioritised age group.

The most popular product categories tend to have low customer ratings, while high-rated products are less popular — suggesting insufficient promotional activity.
Only **6%** of products meet customer quality expectations (rating > 4.0), highlighting a broader challenge in maintaining consistent quality.

The machine-learning model confirmed that **customer engagement and satisfaction** are the primary drivers of product popularity, rather than pricing strategies.

5. Recommendations

- **Improve product quality** to better meet customer expectations.
- **Encourage customers to leave more reviews** to strengthen engagement.
- **Promote high-rated but low-visibility products** through marketing campaigns.
- Focus strategic investments on **product design, durability, and customer satisfaction**, as these factors have a stronger impact on popularity and sales than pricing alone.