Case Study 2: Story Storytelling with Basic Visualization

DAT-230 Data Visualization & Storytelling with AI

Instructor: Dr. Vikas Thammanna Gowda Semester: Fall 2025
Contact: `vthammannagowda@champlain.edu` Office Location: West Hall 100
Office Hours: TBD

Case Study: Used Car Sales Analysis
Assigned Date: 09/11/2025
Due Date: 09/21/2025

1 Overview:

Students will work with a used car sales dataset (features include: price, brand, model, year, mileage, body_type, engine_type, fuel_type, transmission, location, seller_type, vehicle_condition, number_of_owners, listing_age) to explore pricing dynamics, surface bias, and develop value segmentation using visualization and simple modeling. AI/LLM tools (e.g., ChatGPT) may be used as collaborators; all usage must be documented.

1.1 Learning Goals:

- Demonstrate relationships between price and covariates (mileage, age, brand, etc.) via appropriate visualizations.
- Segment vehicles to identify "value" regions and outliers.
- Detect and explain listing biases (e.g., condition vs. price across brands).
- Control for confounders (age, mileage) to estimate brand price premiums.
- Compare transmission types on depreciation and retained value.
- Practice responsible AI-assisted analysis: prompt design, result verification, and reflection.

2 Student Activities

2.1 Visualization Questions

1. Show distribution of prices overall and by brand. Are some brands consistently pricier? Use violin/box plots.
2. Correlate price with mileage and age; visualize with scatter plots plus smoothing (e.g., `geom_smooth()`), color-coded by body type.
3. Compare average price across body types and engine types; investigate whether the SUV vs. sedan price gap holds across fuel types.
4. Build a faceted plot: price vs. mileage stratified by brand and transmission type.
5. Identify outliers: expensive cars with high mileage or cheap cars in excellent condition.
6. Compute and visualize price premium for luxury brands while controlling for age and mileage (e.g., residuals from a regression).
7. Analyze whether automatic transmission cars retain price better than manuals—visualize depreciation curves.
8. *Optional advanced:* Define a "value score" combining condition, mileage, and price; cluster vehicles and visualize cluster characteristics with annotated summaries.

2.2 LLM Prompt Examples

- "Given this used car sales dataset, suggest three `ggplot2` visualizations to compare how price varies with mileage across body types, and generate the R code for the top suggestion."
- "Review this scatter plot of price vs. age colored by brand. List two possible misleading aspects and provide corrected `ggplot2` code with improved encoding and labeling."
- "Write an `rmarkdown` paragraph summarizing whether SUVs command higher prices than sedans after adjusting for mileage and age, including a supporting visualization."
- "Generate a `dplyr` pipeline that computes average price premium per brand after regressing price on mileage and age; explain each step."

2.3 Collaboration and Academic Integrity

- Students may discuss approaches and give peer feedback, but code submissions and reflections must be their own.
- AI assistance is allowed and expected, but must be transparently documented. Copy-pasting LLM output without comprehension or attribution counts as academic dishonesty.
- Collaboration on capstone is individual unless group option is explicitly approved; in group cases, responsibilities must be delineated.
- Use of LLM's/AI to generate reports is strictly prohibited. Any such use will be considered a violation of academic integrity.
- All work must adhere to Champlain College's academic integrity policy; violations will result in disciplinary action.

3 Grading

3.1 Deliverables:

1. **Code:** in `.R` format. (20%)
2. **Report:** a `.pdf` format document containing: (50%)
   Note: You will be provided with a `CaseStudy1_Report.doc` (word document) template for the report with instructions.
   - At least **5 visualizations** with captions and interpretations.
   - Narrative explaining visualizations and insights.
   - Discuss any biases, confounders, and outliers.
3. **Prompt log:** a `.pdf` format document containing: (30%)
   Note: You will be provided with a `CaseStudy1_PromptLog.doc` (word document) template for the prompt log with instructions. You can use screenshots or text where ever you see fit to document your prompts.
   - Summary of LLM prompts for any **2 visualizations** used and their outcomes.
   - Include versions, and any modifications made to LLM outputs.
   - Document how each prompt informed your analysis or visualizations.
   - Be specific about how you verified or corrected LLM outputs.
   - Explain how the use of LLM improved your learning.
   - Explain how the you verified or corrected it.

3.2 Rubric Summary

Note: The following rubric assumes that all three deliverables (code, report, and prompt log) are properly aligned—i.e., the visualizations discussed in the report correspond to the code and the prompts in the log, and there are no mismatches in content or intent. Misalignment (e.g., visuals in the report not reproducible from the code or undocumented prompt-driven changes) will be reflected in the relevant rubric categories and may result in reduced credit.

3.2.1 Code Expectations (20%)

can be assessed qualitatively on:

- Correctness and reproducibility of analysis (e.g., scripts run end-to-end to regenerate key figures).
- Code organization, naming, and use of comments to explain non-obvious steps.
- Appropriate use of tidyverse/ggplot2 idioms and any modeling (e.g., residual calculation) for price premium or similar.

3.2.2 Report Breakdown (50% total)

Table 1: Report Rubric

Criterion	**Points**
At least 5 complete visualizations with clear captions	25
Interpretation of each visualization (insights drawn, relevance to objectives)	25
Narrative coherence connecting visuals to the case study and research questions	15
Discussions identification and implications	15
Clarity, formatting, and adherence to provided template	20
Total	**100**

3.2.3 Prompt Log Breakdown (30% total)

Prompt Log Rubric

Criterion	**Points**
Summary of LLM prompts for at least two visualizations	40
Versions and modifications of model outputs documented	10
Explanation of how each prompt informed analysis or visualization choices	15
Reflection on how LLM use improved learning and understanding	15
Clarity, formatting, and adherence to provided template	20
Total	**30**