Project Deliverable 1: Amazon Customer Behavior dataset analysis

Mausam Shrestha

MSCS-634, Advanced Big Data and Data Mining

Department of Computer Information Sciences, University of the Cumberland's

Nov 2nd, 2025

Source: https://www.kaggle.com/datasets/swathiunnikrishnan/amazon-consumer-behaviour-dataset

GitHub: https://github.com/shrestha-mausam/MSCS_634_ProjectDeliverable_1

1. Introduction

1.1 Project Overview

This report presents a comprehensive analysis of the Amazon Consumer Behavior Dataset, which contains consumer behavior patterns and shopping preferences from Amazon customers. The primary objective of this project is to clean, explore, and analyze the dataset to identify factors influencing consumer behavior, shopping satisfaction, purchase frequency, and review behavior.

The analysis serves as the foundation for future predictive modeling tasks focused on understanding and predicting consumer behavior patterns. By examining relationships between demographic variables, shopping preferences, and satisfaction metrics, this study aims to provide actionable insights for improving customer experience and business strategies.

1.2 Dataset Description

The Amazon Consumer Behavior Dataset is a publicly available dataset obtained from Kaggle, containing survey responses from Amazon customers. The dataset includes:

- **Sample Size:** 602 records
- **Features:** 23 attributes
- **Memory Usage:** Approximately 0.68 MB
- **Data Types:**
  - 5 numerical variables (age, satisfaction scores, importance ratings)
  - 18 categorical variables (purchase patterns, preferences, demographics)

1.3 Key Variables

The dataset encompasses several categories of variables:

Demographics: - Age - Gender

Purchase Behavior: - Purchase Frequency - Purchase Categories - Browsing Frequency - Cart Completion Frequency - Cart Abandonment Factors

Shopping Experience: - Shopping Satisfaction (1-5 scale) - Service Appreciation - Improvement Areas - Product Search Method

Review Behavior: - Review Left (Yes/No) - Review Reliability - Review Helpfulness - Customer Reviews Importance

Recommendation Systems: - Personalized Recommendation Frequency - Recommendation Helpfulness - Rating Accuracy

1.4 Project Objectives

The main objectives of this analysis are:

1. **Data Quality Assurance:** Identify and address data quality issues including missing values, duplicates, inconsistencies, and outliers

2. **Exploratory Analysis:** Understand data distributions, relationships between variables, and behavioral patterns
3. **Pattern Identification:** Discover key insights about consumer behavior and factors influencing satisfaction
4. **Foundation for Modeling:** Prepare the dataset and document findings to guide future predictive modeling efforts

---

2. Data Preparation Steps

2.1 Data Loading

The dataset was downloaded using the KaggleHub API, requiring authentication via API credentials stored in ~/.kaggle/kaggle.json. The dataset file "Amazon Customer Behavior Survey.csv" was loaded into a pandas DataFrame for analysis.

2.2 Initial Data Inspection

A comprehensive initial inspection revealed:

- **Shape:** 602 rows × 23 columns
- **Missing Values:** 2 missing values (0.33%) in the Product_Search_Method column
- **Duplicate Rows:** No duplicate rows found
- **Column Name Issues:** Two columns had trailing spaces:
  - 'Personalized_Recommendation_Frequency ' (trailing space)
  - 'Rating_Accuracy ' (trailing space)

2.3 Handling Missing Values

Challenge Identified: - 2 missing values in the Product_Search_Method column (0.33% of data)

Solution Applied: - **Mode Imputation:** Since the missing values constituted a minimal percentage and the variable is categorical, missing values were filled with the mode (most frequent value: "categories")

Rationale: - Minimal impact on data distribution (less than 1% missing) - Mode imputation preserves categorical distribution patterns - Simple and effective approach for categorical variables with low missing rates

Result: - All missing values successfully handled - No data loss occurred

2.4 Removing Duplicates and Correcting Inconsistencies

2.4.1 Duplicate Detection

- **Method:** Used pandas duplicated() method to check for exact duplicate rows
- **Result:** No duplicate rows found in the dataset

2.4.2 Column Name Standardization

- **Issue Identified:** Columns with trailing spaces causing potential KeyError issues

  - 'Personalized_Recommendation_Frequency ' → Renamed to 'Personalized_Recommendation_Frequency_Num'
  - 'Rating_Accuracy ' → Renamed to 'Rating_Accuracy'
- **Action:** Applied systematic column name cleaning

2.4.3 Categorical Value Cleaning

- **Issue Identified:** Trailing/leading spaces in categorical values
  - Found in: Service_Appreciation and Improvement_Areas columns
- **Solution:** Applied .strip() method to all object-type columns to remove whitespace inconsistencies
- **Result:** Standardized categorical values across the dataset

2.5 Identifying and Addressing Noisy Data

2.5.1 Outlier Detection

Using the Interquartile Range (IQR) method, outliers were identified in numerical columns:

Variable	Outliers	Percentage	Normal Range	Actual Range
Age	20	3.32%	[3.5, 55.5]	[3, 67]
Personalized Recommendation Frequency	35	5.81%	[0.5, 4.5]	[1, 5]
Rating Accuracy	21	3.49%	[0.5, 4.5]	[1, 5]
Shopping Satisfaction	17	2.82%	[0.5, 4.5]	[1, 5]
Customer Reviews Importance	0	0.00%	N/A	[1, 5]

2.5.2 Age Validation

Potentially Unrealistic Ages Identified: - Age 3 (1 occurrence) - Possibly a parent responding on behalf of a child - Age 12 (1 occurrence) - Lower age boundary case - Ages above 55 (various occurrences up to 67)

Decision Rationale: - Preserved all age values as they may represent valid survey responses - Age 3 likely represents proxy responses (parent answering for child) - Older ages (55-67) are plausible for online shopping demographics - Documented for future reference if modeling requires age constraints

2.5.3 Categorical Inconsistencies Check

- Checked for case variations and typos across categorical columns
- **Result:** No major inconsistencies found after initial cleaning operations

2.6 Data Quality Summary

After completing all data preparation steps:

- ✓ **Missing Values:** 0 (all handled)

- ✓ **Duplicates:** 0 (none found)
- ✓ **Column Names:** All standardized
- ✓ **Categorical Values:** All cleaned of whitespace issues
- ✓ **Outliers:** Documented and preserved (valid survey responses)
- ✓ **Final Dataset Shape:** 602 rows × 23 columns (no data loss)

---

3. Modeling Details

3.1 Current Stage

Note: At this stage of the project (Deliverable 1), the focus has been on data cleaning and exploratory data analysis. Predictive modeling has not yet been implemented. This section documents the modeling approach planned for future phases.

3.2 Target Variable Selection

Based on the exploratory analysis, three potential target variables have been identified:

Primary Target:

- **Shopping_Satisfaction** (Ordinal, 1-5 scale)
  - Represents overall customer satisfaction
  - Moderate correlation with multiple features
  - Suitable for ordinal regression

Secondary Targets:

- **Review_Left** (Binary: Yes/No)
  - Indicates whether customers leave product reviews
  - Suitable for binary classification
  - Business value: Understanding review engagement
- **Purchase_Frequency** (Categorical)
  - Predicts customer purchase behavior frequency
  - Suitable for multi-class classification
  - Business value: Customer segmentation and retention

3.3 Feature Engineering Recommendations

The following feature engineering steps are recommended for modeling:

3.3.1 Age Group Segmentation

- **Status:** Already implemented in EDA
- **Bins:** 18-25, 26-35, 36-45, 45+
- **Rationale:** Categorical representation may improve model performance

3.3.2 Purchase Categories Feature Extraction

- Extract count of categories purchased

- Create binary indicators for presence of specific categories
- Handle multi-value categorical strings (e.g., "Beauty and Personal Care;Clothing and Fashion")

3.3.3 Composite Scores

- Combine related review metrics into composite scores
- Create aggregated satisfaction indicators

3.3.4 Interaction Terms

- Age × Gender interactions
- Purchase_Frequency × Satisfaction interactions
- Recommendation Frequency × Satisfaction interactions

3.3.5 Temporal Features (if relevant)

- Extract features from Timestamp column (day of week, time of day, season)
- Or remove if not relevant to model objectives

3.4 Modeling Approaches Recommended

3.4.1 For Shopping_Satisfaction (Ordinal Target)

- **Ordinal Regression Models:**
  - Ordered Logit Model
  - Ordinal Random Forest
  - Neural Network with ordinal output layer
- **Rationale:** Respects the 1-5 ordering of satisfaction levels

3.4.2 For Review_Left (Binary Target)

- **Classification Models:**
  - Logistic Regression (baseline)
  - Random Forest Classifier
  - Gradient Boosting (XGBoost, LightGBM)
  - Support Vector Machine
- **Rationale:** Standard binary classification problem

3.4.3 For Purchase_Frequency (Multi-class Target)

- **Classification Models:**
  - Random Forest Classifier
  - Gradient Boosting Classifier
  - Naive Bayes (for baseline)

3.4.4 Ensemble Methods

- **Rationale:** Mixed data types (numerical + categorical) benefit from ensemble approaches
- **Options:** Voting classifiers, stacking, or boosting methods

3.5 Preprocessing Steps Needed

1. **Categorical Encoding:**

- One-hot encoding for nominal variables (Gender, Product_Search_Method)
- Ordinal encoding for frequency-based categoricals (Purchase_Frequency, Browsing_Frequency)
- Label encoding for ordinal scales (if treating as categorical)

2. **Numerical Scaling:**
   - StandardScaler or MinMaxScaler (if using distance-based algorithms)
   - Not required for tree-based models
3. **Feature Selection:**
   - Remove 'Timestamp' or convert to temporal features
   - Consider feature importance from exploratory analysis
4. **Class Imbalance Handling:**
   - Check Review_Left distribution for imbalance
   - Apply SMOTE, undersampling, or class weights if needed

3.6 Validation Strategy

1. **Train-Test Split:**
   - 80-20 or 70-30 split maintaining distribution of key categorical variables
   - Stratified splitting for categorical targets
2. **Cross-Validation:**
   - Stratified k-fold cross-validation (k=5 or k=10)
   - Ensures balanced representation across folds
3. **Overfitting Prevention:**
   - Monitor train vs. validation performance
   - Use early stopping for iterative models
   - Regularization techniques
   - Consider simpler models given sample size (602 rows)

---

4. Evaluation

4.1 Data Quality Evaluation

4.1.1 Completeness

- **Before Cleaning:** 2 missing values (0.33%)
- **After Cleaning:** 0 missing values (100% complete)
- **Assessment:** Excellent data completeness

4.1.2 Consistency

- **Column Names:** Standardized (trailing spaces removed)
- **Categorical Values:** Whitespace inconsistencies eliminated
- **Assessment:** High consistency achieved

4.1.3 Accuracy

- **Outliers:** Documented and validated as potential valid responses

- **Age Validation:** Unusual ages preserved with rationale documented
- **Assessment:** Data accuracy validated within context

4.1.4 Validity

- All numerical ranges within expected bounds
- Categorical values align with survey structure
- **Assessment:** Data validity confirmed

4.2 Exploratory Analysis Evaluation

4.2.1 Distribution Analysis

Numerical Variables: - **Age:** Right-skewed distribution (median: 26, range: 3-67) - **Shopping Satisfaction:** Clustered around moderate levels (mean: 2.46, median: 2.0) - **Rating Accuracy:** Similar clustering (mean: 2.67, median: 3.0) - **Customer Reviews Importance:** Bimodal distribution (peaks at 1 and 3)

Categorical Variables: - **Gender:** Female-dominated (58.5%) - **Purchase Frequency:** "Few times a month" most common (34%) - **Browsing Frequency:** "Few times a week" most common (41%) - **Review Behavior:** Approximately balanced between Yes/No

4.2.2 Relationship Analysis

Key Correlations Identified:

Variable Pair	Correlation	Strength	Interpretation
Shopping Satisfaction ↔ Rating Accuracy	0.514	Moderate	Higher rating accuracy associated with higher satisfaction
Shopping Satisfaction ↔ Recommendation Frequency	0.438	Moderate	More recommendations correlate with higher satisfaction
Reviews Importance ↔ Satisfaction	0.402	Moderate	Reviews importance linked to overall satisfaction
Recommendation Frequency ↔ Rating Accuracy	0.438	Moderate	Recommendation systems and rating accuracy are related

Behavioral Patterns Discovered: - Purchase frequency positively correlates with review engagement - Personalized recommendations show varying impact across frequency levels - Age groups exhibit distinct behavioral patterns - Younger consumers (18-25) show different purchase patterns

4.3 Statistical Summary

Numerical Variables Summary:

Variable	Mean	Median	Std Dev	Min	Max
Age	30.79	26.00	10.19	3	67
Customer Reviews Importance	2.48	3.00	1.19	1	5

Personalized Recommendation Frequency	2.70	3.00	1.04	1	5
Rating Accuracy	2.67	3.00	0.90	1	5
Shopping Satisfaction	2.46	2.00	1.01	1	5

4.4 Model Readiness Assessment

Strengths: - ✓ Clean, complete dataset - ✓ Well-documented data quality issues - ✓ Clear target variables identified - ✓ Feature relationships understood - ✓ Preprocessing requirements documented

Limitations: - Sample size (602 rows) may limit complex model complexity - Mixed data types require careful encoding strategy - Some categorical variables have many unique values - Ordinal nature of satisfaction scales needs appropriate modeling approach

Readiness Score: Ready for Modeling (with appropriate techniques for sample size)

---

5. Key Insights

5.1 Demographic Insights

1. **Age Distribution:**
   - Primary consumer base: 23-36 years old (median: 26)
   - Younger demographic dominance suggests focus on digital-native consumers
   - Age outliers (3, 67) represent edge cases but valid responses
2. **Gender Distribution:**
   - Female consumers represent 58.5% of the sample
   - Gender-based behavioral differences observed in purchase patterns
   - Consideration needed for gender-balanced marketing strategies

5.2 Purchase Behavior Insights

1. **Purchase Frequency:**
   - Most common pattern: “Few times a month” (34%)
   - Suggests regular but not excessive shopping behavior
   - Opportunities for retention strategies targeting less frequent shoppers
2. **Browsing vs. Purchasing:**
   - Browsing frequency (41% browse “Few times a week”) exceeds purchase frequency
   - Indicates window shopping behavior
   - Opportunity: Convert browsers to purchasers through targeted strategies
3. **Product Search Methods:**
   - Categories (37%) and Keywords (36%) are nearly equal
   - Users comfortable with both navigation methods
   - Recommendation: Maintain both search capabilities

5.3 Satisfaction and Experience Insights

1. **Satisfaction Levels:**
   - Clustering around moderate satisfaction (levels 2-3 on 5-point scale)
   - Mean satisfaction: 2.46 suggests room for improvement
   - Opportunity: Identify drivers of higher satisfaction
2. **Service Appreciation:**
   - Top factors: "Competitive prices" and "Product recommendations"
   - Price sensitivity remains important
   - Personalization valued by consumers
3. **Improvement Areas:**
   - Top concerns: "Product quality and accuracy" and "Customer service responsiveness"
   - Actionable feedback for business improvements
   - Quality assurance and customer service are priority areas

5.4 Behavioral Relationship Insights

1. **Satisfaction Drivers:**
   - **Rating Accuracy (r=0.514):** Strongest predictor of satisfaction
     - Implication: Accurate product ratings critical for customer satisfaction
   - **Personalized Recommendations (r=0.438):** Strong positive relationship
     - Implication: Effective recommendation systems boost satisfaction
   - **Reviews Importance (r=0.402):** Moderate positive relationship
     - Implication: Review systems contribute to satisfaction
2. **Review Behavior:**
   - Frequent purchasers more likely to leave reviews
   - Suggests engaged customers provide more feedback
   - Opportunity: Encourage reviews from occasional purchasers
3. **Age Group Differences:**
   - Younger consumers (18-25) show distinct purchase patterns
   - Age-based segmentation may improve targeting
   - Personalization strategies should consider age groups

5.5 Business Implications

1. **Recommendation Systems:**
   - Strong correlation with satisfaction suggests investment in personalization is valuable
   - Focus on improving recommendation accuracy and relevance
2. **Product Quality:**
   - Top improvement area indicates need for quality assurance
   - May impact satisfaction scores and customer retention
3. **Customer Service:**
   - Responsiveness identified as improvement area

- May correlate with satisfaction and retention

4. **Review Systems:**
   - Reviews important to consumers and satisfaction
   - Encourage review engagement, especially from occasional purchasers

---

6. Ethical Considerations

6.1 Data Privacy and Consent

Considerations: - The dataset is publicly available on Kaggle with appropriate licensing - No personally identifiable information (PII) beyond demographics in the dataset - Analysis uses aggregated patterns rather than individual-level profiling

Recommendations: - Continue to ensure no PII is used in modeling or reporting - Respect data source licensing and attribution requirements - If deploying models, ensure compliance with data privacy regulations (GDPL, CCPA, etc.)

6.2 Bias and Fairness

6.2.1 Demographic Bias

Identified Concerns: - Gender imbalance: 58.5% female, potentially underrepresenting male perspectives - Age distribution skews toward younger demographics (median: 26) - Potential geographic bias (dataset source and collection method not fully documented)

Mitigation Strategies: - Acknowledge demographic limitations in any conclusions - Avoid making broad generalizations beyond the dataset demographics - Consider stratified analysis by demographic groups - If deploying models, monitor for demographic bias in predictions

6.2.2 Sampling Bias

Concerns: - Sample size (602) may not represent all Amazon customer segments - Self-selected survey respondents may introduce participation bias - Voluntary survey responses may not reflect true population behavior

Mitigation: - Clearly state sample limitations in all reporting - Avoid extrapolating beyond the dataset characteristics - Consider the self-selection bias in interpretation

6.3 Use of Predictive Models

6.3.1 Ethical Model Deployment

Considerations: - Predictive models should not be used for discriminatory practices - Avoid models that could disadvantage certain demographic groups - Ensure transparency in model decisions affecting individuals

Recommendations: - Implement fairness metrics in model evaluation - Regular audits for demographic parity - Transparency in model limitations and appropriate use cases

6.3.2 Manipulation Concerns

Risks: - Models predicting purchase frequency could be used for manipulative marketing - Satisfaction prediction models might be used to exploit vulnerable consumers - Personalization could lead to price discrimination

Mitigation: - Use models to improve customer experience, not manipulate behavior - Maintain ethical marketing practices - Ensure personalization benefits both customers and business fairly

6.4 Data Quality and Reliability

6.4.1 Survey Data Limitations

Concerns: - Self-reported data may contain inaccuracies or social desirability bias - Survey responses may not reflect actual behavior - Outliers (e.g., age 3) require careful interpretation

Approach: - Documented data quality issues transparently - Preserved outliers with rationale rather than removing arbitrarily - Acknowledged limitations in all conclusions

6.5 Recommendations for Ethical Practice

1. **Transparency:**
   - Clearly document data sources, limitations, and methodology
   - Acknowledge demographic biases and sampling limitations
   - Report model performance metrics honestly
2. **Fairness:**
   - Monitor for demographic bias in models
   - Ensure models do not perpetuate or amplify existing inequalities
   - Implement fairness metrics in evaluation
3. **Privacy:**
   - Ensure no PII is used or exposed
   - Follow data protection regulations
   - Respect user privacy in any deployed applications
4. **Responsible Use:**
   - Use insights to improve customer experience, not manipulate behavior
   - Consider societal impact of recommendations and targeting
   - Maintain ethical standards in business applications

---

7. Recommendations

7.1 Data Collection Recommendations

1. **Increase Sample Size:**
   - Current sample (602) limits complex modeling
   - Recommend expanding to 2000+ records for robust model development
   - Improved statistical power for detecting relationships
2. **Demographic Balance:**

- Collect more balanced gender representation
- Include broader age range representation
- Consider geographic diversity if relevant

3. **Data Collection Methods:**
   - Supplement survey data with actual behavioral data if possible
   - Validate self-reported data against actual behavior
   - Consider longitudinal data collection for temporal analysis

7.2 Feature Engineering Recommendations

1. **Purchase Categories:**
   - Extract category count as numerical feature
   - Create binary indicators for specific popular categories
   - Consider category embeddings for high-dimensional categorical data
2. **Temporal Features:**
   - Extract day of week, time of day from Timestamp if relevant
   - Consider seasonality if data spans multiple months
   - Remove Timestamp if temporal patterns not relevant
3. **Interaction Terms:**
   - Test Age × Gender interactions
   - Explore Purchase_Frequency × Recommendation interactions
   - Consider Satisfaction × Service_Appreciation interactions
4. **Composite Metrics:**
   - Create review engagement score combining multiple review-related features
   - Develop overall satisfaction composite from related metrics
   - Build recommendation system effectiveness score

7.3 Modeling Recommendations

1. **Model Selection:**
   - Start with simpler models (logistic regression, random forest) due to sample size
   - Use ordinal regression for satisfaction prediction
   - Consider ensemble methods for robustness
2. **Validation Approach:**
   - Implement stratified k-fold cross-validation
   - Use separate validation set for final model evaluation
   - Monitor for overfitting given limited sample size
3. **Evaluation Metrics:**
   - For ordinal satisfaction: Use appropriate ordinal metrics (ordinal accuracy, mean absolute error)
   - For binary classification: Use precision, recall, F1-score, ROC-AUC
   - For multi-class: Use confusion matrix, per-class metrics

7.4 Business Recommendations

7.4.1 Improve Customer Satisfaction

1. **Rating Accuracy (Highest Correlation with Satisfaction):**
   - Invest in rating quality assurance systems
   - Implement mechanisms to detect and prevent rating manipulation
   - Improve product description accuracy to align with ratings
2. **Personalized Recommendations (Second Highest Correlation):**
   - Enhance recommendation algorithm accuracy
   - Improve recommendation relevance
   - Test different recommendation frequencies to optimize satisfaction
3. **Review Systems:**
   - Encourage more customers to leave reviews (especially occasional purchasers)
   - Improve review quality and helpfulness
   - Make reviews more accessible and useful to shoppers

7.4.2 Address Improvement Areas

1. **Product Quality and Accuracy:**
   - Strengthen quality assurance processes
   - Improve product descriptions and images
   - Enhance vendor/seller quality standards
2. **Customer Service Responsiveness:**
   - Reduce response times
   - Improve customer service availability
   - Enhance service quality training
3. **Packaging Waste:**
   - Implement sustainable packaging initiatives
   - Offer eco-friendly packaging options
   - Communicate sustainability efforts to customers

7.4.3 Marketing and Personalization

1. **Segmentation Strategy:**
   - Develop age-based segmentation strategies
   - Create targeted campaigns for different purchase frequency groups
   - Personalize based on browsing behavior
2. **Engagement:**
   - Convert frequent browsers to purchasers
   - Encourage review engagement from all customer segments
   - Optimize recommendation frequency for different customer types

7.5 Technical Recommendations

1. **Data Infrastructure:**
   - Set up automated data quality monitoring

   - Implement data validation pipelines
   - Create standardized data cleaning procedures
2. **Model Deployment:**
   - Develop model monitoring and retraining pipelines
   - Implement A/B testing frameworks for model improvements
   - Create model explainability tools for business stakeholders
3. **Documentation:**
   - Maintain detailed data dictionaries
   - Document all preprocessing steps
   - Create reproducible analysis workflows

---

8. References

8.1 Dataset Source

Swathi Unnikrishnan. (2023). *Amazon Consumer Behaviour Dataset*. Kaggle. https://www.kaggle.com/datasets/swathiunnikrishnan/amazon-consumer-behaviour-dataset

8.2 Software and Libraries

- **Python 3.13:** Programming language
- **Pandas 2.3.3:** Data manipulation and analysis
  - McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51-56).
- **NumPy 2.3.4:** Numerical computations
  - Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357-362.
- **Matplotlib 3.10.7:** Data visualization
  - Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95.
- **Seaborn 0.13.2:** Statistical data visualization
  - Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021.
- **KaggleHub 0.3.13:** Dataset download utility
  - KaggleHub documentation: https://github.com/Kaggle/kagglehub

8.3 Methodological References

- **Data Cleaning and Preprocessing:**
  - Rahman, M. M., & Islam, M. Z. (2013). Missing value imputation using decision trees and decision forests. *International Journal of Information Technology and Computer Science*, 5(9), 72-79.
- **Outlier Detection:**
  - Aggarwal, C. C. (2017). *Outlier analysis*. Springer.
- **Exploratory Data Analysis:**

  - Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- **Ordinal Regression:**
  - Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). John Wiley & Sons.
- **Feature Engineering:**
  - Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: Principles and techniques for data scientists*. O'Reilly Media.

8.4 Ethical Guidelines

- **Data Privacy:**
  - General Data Protection Regulation (GDPR). (2018). Regulation (EU) 2016/679.
  - California Consumer Privacy Act (CCPA). (2018). California Civil Code § 1798.100-1798.199.
- **Fairness in Machine Learning:**
  - Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org.
  - Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5), 82-89.

8.5 Business Intelligence References

- **Customer Satisfaction Analysis:**
  - Anderson, E. W., & Sullivan, M. W. (1993). The antecedents and consequences of customer satisfaction for firms. *Marketing Science*, 12(2), 125-143.
- **Recommendation Systems:**
  - Ricci, F., Rokach, L., & Shapira, B. (2015). *Recommender systems handbook* (2nd ed.). Springer.
- **Consumer Behavior:**
  - Solomon, M. R. (2014). *Consumer behavior: Buying, having, and being* (11th ed.). Pearson.