Applied Regression Methods Project

Mohamed Ibrahim Abdelmagid

November 2024

Statement of Purpose

What Questions Can Be Answered by the Analysis of the Data?

The analysis of the data will help answer several key questions related to the impact of weather on consumer mobility and ATM withdrawal patterns:

1. **How does weather variability influence consumer mobility?**

   By examining the relationship between weather conditions and transportation data, we can determine whether adverse weather (e.g., heavy rain, snow, or extreme temperatures) leads to a decrease in consumer mobility, as suggested by previous studies [1]. This analysis is crucial

for understanding how environmental factors impact consumer movement, which could, in turn, inform urban planning and transportation system design.

2. **Can weather data be used to predict financial transaction patterns?**

   By correlating weather data with transaction trends, this study may develop predictive models that forecast ATM withdrawals based on upcoming weather conditions. This can have practical implications for financial institutions, helping them plan for periods of higher or lower transaction volumes due to weather. Such models could also assist in optimizing ATM placement and operational strategies in weather-sensitive areas.

These questions will guide the study's data analysis and provide insights into the broader effects of weather on consumer behavior and economic activities, especially in relation to mobility and financial transactions.

Literature Review

Understanding the factors influencing consumer mobility and ATM withdrawals is critical for designing efficient financial and urban systems. Weather conditions significantly impact human behavior and mobility patterns, as highlighted in prior studies. For instance, Rudloff et al. (2015) analyzed travel behavior in Vienna and noted that weather variables such as rainfall and temperature could alter individuals' mode of transportation and trip frequency. These findings underscore the need to incorporate weather variability into models of mobility and economic activity, especially in urban settings [1].

Moreover, seasonal variations in weather conditions influence consumption patterns. Sandqvist and Siliverstovs (2021) presented an intertemporal consumption model demonstrating how weather conditions act as utility modifiers. They revealed that unusual weather events could shift seasonal consumption, thereby affecting economic activities like retail and ATM withdrawals. This relationship highlights the broader economic implications of weather changes on consumer behavior [2].

Specific to financial transactions, studies suggest that adverse weather conditions reduce consumer outdoor activity, potentially lowering ATM withdrawals. Conversely, favorable weather might encourage mobility, increasing transactions. A comprehensive framework for modeling such behavior must consider the interplay between socio-economic factors, weather data, and mobility patterns, as evidenced by Sandqvist et al.'s theoretical formulations and empirical studies [2].

In conclusion, existing literature provides strong evidence of the impact of weather on consumer behavior and mobility. However, further research is needed to quantitatively link these insights with real-time ATM transaction data to develop predictive models tailored to specific urban contexts.

Description of the Data

A Complete Source of the Data

Here is a link to the Kaggle dataset: **Click Here**.

The data originates from two main sources:

- **Spar Nord**: Information related to ATM transactions, such as location, currency, card type, and ATM service.
- **OpenWeatherMap**: Weather data, including temperature, pressure, humidity, wind speed, and other weather-related variables (https://openweathermap.org/).

What are the Observations?

The observations are **individual ATM withdrawal transactions**, each including data about the transaction, ATM details, and corresponding weather conditions.

Definition of Variables, Including Type and Unit of Measurement

Variable Name	Description	Type	Values / Unit
`year`	The year the withdrawal was made	Quantitative	Integer
`month`	The month the withdrawal was made	Categorical	Text (January–December)
`day`	The day of the month the withdrawal was made	Quantitative	Integer (1–31)
`weekday`	The day of the week the withdrawal was made	Categorical	Text (Sunday–Saturday)

(Continued on next page)

(Continued from previous page)

Variable Name	Description	Type	Values / Unit
hour	The hour of the day the withdrawal was made	Quantitative	Integer (1–24)
atm_status	Status of the ATM (active or inactive)	Categorical	Text (Active, Inactive)
atm_id	Unique ID number for each ATM	Categorical	Text (1–113)
atm_manufacturer	Manufacturer of the ATM	Categorical	Text (e.g., NCR, Diebold Nixdorf)
atm_location	Location (city or description)	Categorical	Text (113 unique strings)
atm_streetname	Street name where the ATM is located	Categorical	Text (83 unique strings)
atm_street_number	Street number where the ATM is located	Quantitative	Integer (1–452)
atm_zipcode	Zip code of the ATM location	Quantitative	Integer (1550–9990)
atm_lat	Latitude of the ATM location (WGS84 coordinates)	Quantitative	Numeric (55.06–57.72)
atm_lon	Longitude of the ATM location (WGS84 coordinates)	Quantitative	Numeric (8.408–12.612)
currency	Currency used in the withdrawal	Categorical	Text (DKK, EUR, GBP, USD)

(Continued on next page)

(Continued from previous page)

Variable Name	Description	Type	Values / Unit
`card_type`	Type of card used for withdrawal	Categorical	Text (12 unique strings)
`service`	Type of ATM service used	Categorical	Text (Withdrawal)
`message_code`	Error code during withdrawal	Categorical	Text (or missing if no error)
`message_text`	Error message corresponding to the error code	Categorical	Text (or missing if no error)
`weather_lat`	Latitude of weather measurement location (WGS84 coordinates)	Quantitative	Numeric (55.06–57.72)
`weather_lon`	Longitude of weather measurement location (WGS84 coordinates)	Quantitative	Numeric (8.450–12.614)
`weather_city_id`	ID of the city where weather was measured	Quantitative	Integer
`weather_city_name`	Name of the city where weather was measured	Categorical	Text (53 unique strings)
`temp`	Temperature at the moment of measurement (Kelvin)	Quantitative	Numeric (260.0–302.1)
`pressure`	Atmospheric pressure at sea level (hPa)	Quantitative	Integer (974–1057)
`humidity`	Humidity percentage	Quantitative	Integer (0–174)

(Continued on next page)

(Continued from previous page)

Variable Name	Description	Type	Values / Unit
`wind_speed`	Wind speed (meters per second)	Quantitative	Numeric (0–77)
`wind_deg`	Wind direction (meteorological degrees)	Quantitative	Integer (0–360)
`rain_3h`	Rain volume in the last 3 hours (mm)	Quantitative	Numeric (0.1–25.4)
`clouds_all`	Cloudiness percentage	Quantitative	Integer (0–100)
`weather_id`	Weather condition ID	Quantitative	Integer
`weather_main`	Group of weather parameters	Categorical	Text (e.g., Rain, Clouds, Thunder-storm)
`weather_description`	Detailed description of weather conditions	Categorical	Text (32 unique strings)

Which Variable is Used as the Response?

The **response variable** is the `number of withdrawals per day`.

Hypothesis or Reasoning for Predictor Influence

- **Time Variables (`year`, `month`, `day`, `weekday`, `hour`)**: Likely influence withdrawal trends (e.g., higher activity on weekends or during business hours).

- **Weather-related Variables** (`temp`, `pressure`, `humidity`, `wind_speed`): Extreme weather conditions may reduce withdrawals due to lower mobility.

Data Analysis

The data analysis process involved several key steps, starting with exploring the data, transforming variables if necessary, fitting the initial regression model, and critically evaluating the model as shown in figure 1:

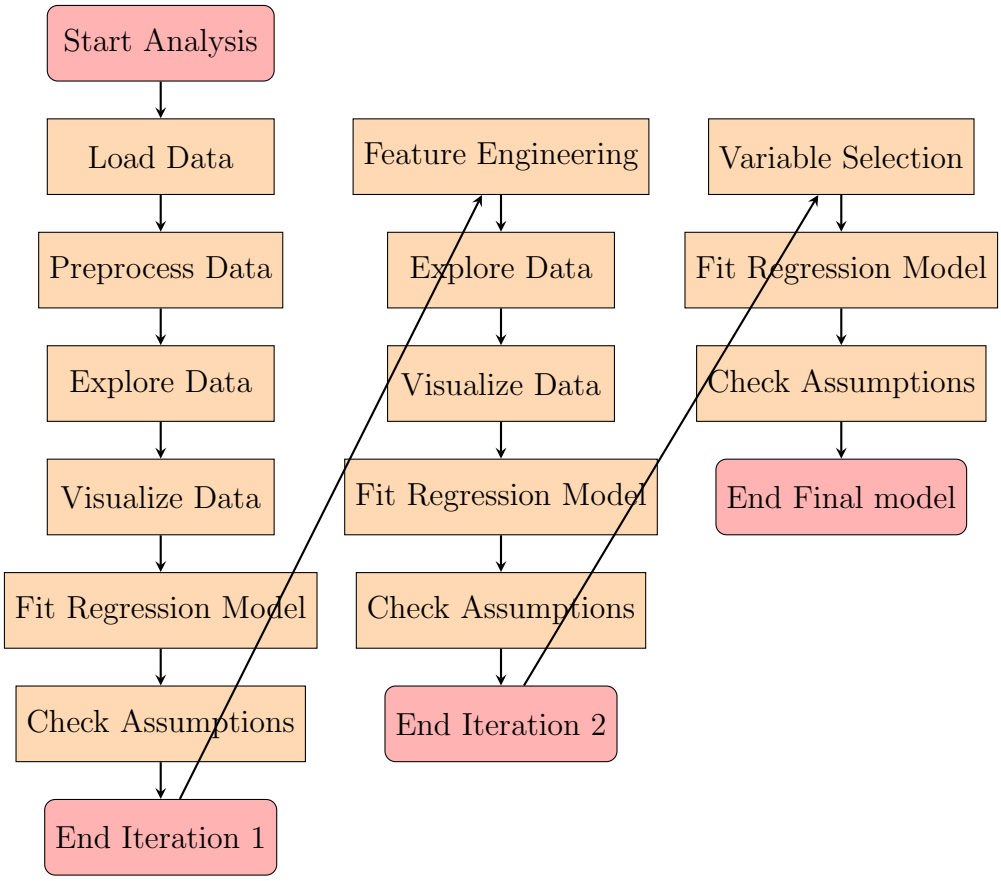

Figure 1: Data Analysis Iterations Workflow

Data preprocessing

The dataset was cleaned by filtering rows for a specific `atm_id`, removing irrelevant columns (e.g., `currency`, `card_type`, atm-related features), and dummy encoding categorical variables while avoiding multicollinearity. Furthermore, the dataset was cleaned further for a specific ATM only for our analysis. The ATM selected was in Næstved, Farimagsvej street.

Graphs Before Fitting a Model

Before fitting the model, various exploratory data analysis (EDA) steps were taken, including visualizing the data through scatter plots and examining the pairwise relationships between variables. There was a correlation between month and temp as they have a correlation of 0.82. However, after examining the graph, it does not have a strong relation as the month variable is nominal. Also, the variance inflation factor was not high in both variables.

1st Iteration

Transformation of Variables

In the 1st iteration, it was noticed the existence of categorical variables. Thus, such variables were encoded to be used in the regression properly. Second, it was noticed in the graphs before fitting the model a correlation between **month** and **temp**. This is reasonable as certain months tend to have higher temperature than others. Second, the dataset was sorted according to day and month

to make sure testing the index plot of residuals is meaningful.

Initial Model Fit to the Data

An initial linear regression model was fitted using the `lm()` function, where the dependent variable `service` was modeled as a function of all other predictors. The final R^2 was 0.35 with an R^2_{adj} of 0.26.

Graphs After Fitting the Model

After fitting the model, diagnostic plots were generated to assess the model's fit and identify potential issues. The following diagnostic plots were created:

- Fitted vs actual values: A linear trend was noticed. However, there are many outliers in the actual values indicating more transformation need to capture the right relation 2.

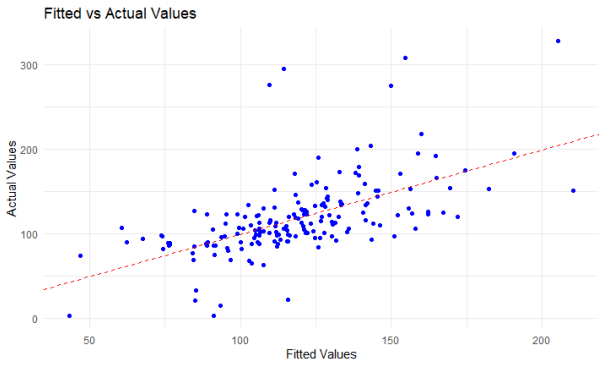

Figure 2: Iteration 1: fitted vs actual

- Residuals vs. Fitted Values plot

– It was random; but there were outliers 3.

Figure 3: Index plot of the residuals

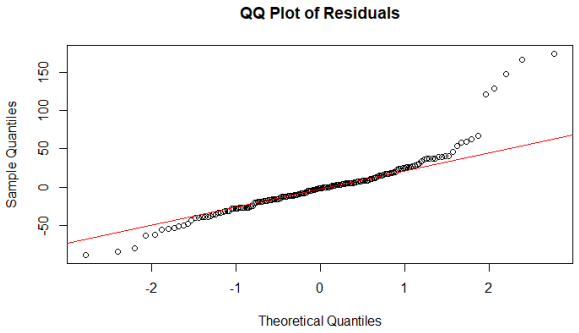

Figure 4: QQ plot iteration 1

- Normal Q-Q plot of residuals
  - It follows the linear trend but there were higher residuals than expected for higher fitted values. This may also be reasoned by the fact that most of the y-values are centered around the mean 4.
- Residuals vs Predictors: Some variables had abnormal residuals such as the clouds_all vs studentized residuals 5.

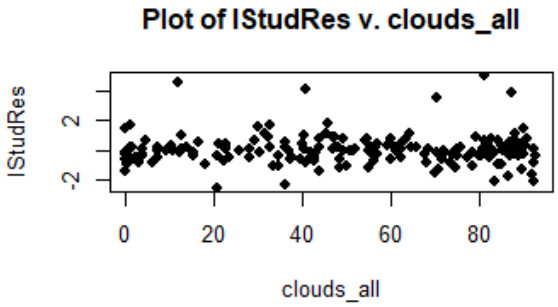

Figure 5: Some points have abnormal residuals than others

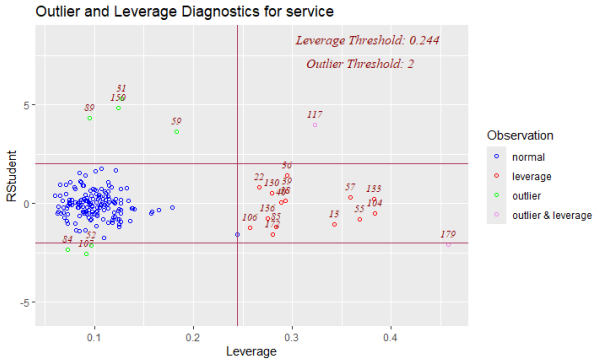

Figure 6: Potential residual plot iteration 1

- Leverage vs. Standardized Residuals plot
  - Few points have high leverage and low residuals but there were a lot of points with high residuals and low leverage 6.

Model Criticism and Reformulation

Upon examining the diagnostic plots, it became clear that a transformation is needed for some of the predictors. Also, the five points in the index-residual plot were validated to be true observations, and thus data transformation is needed.

Iteration 2

Feature Engineering

Several new features were engineered to capture more relevant patterns in the data. Specifically:

- `end_of_month`: This binary variable indicates whether a given day is at the end of the month (e.g., days 28 to 31).
- `weekend`: A binary indicator representing whether the day is a weekend, derived from the weekday columns.
- `end_of_month_and_weekend`: This interaction term captures the overlap between weekends and end-of-month days.
- `lag_1_service`: This feature captures the lagged value of the `service` variable, helping to account for temporal dependencies.
- `service_day_trend`: A quadratic transformation of the `day` variable, centered around the 15th day of the month, to capture the trend in the index plot of the residuals.

Graphs After Fitting the Model

After fitting the model, diagnostic plots were generated to assess the model's fit and identify potential issues. Most of the graphs were similar to the previous iteration; thus, only notable changes will be included:

- Normal Q-Q plot of residuals
  - It still follows the linear trend but the residuals of some of the points have been reduced as shown in 7.

- Leverage vs. Standardized Residuals plot
  - The residuals of some of the points have been reduced as well as the potential 8.

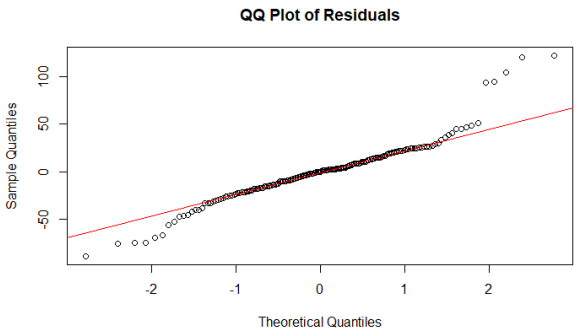

Figure 7: QQ plot iteration 2

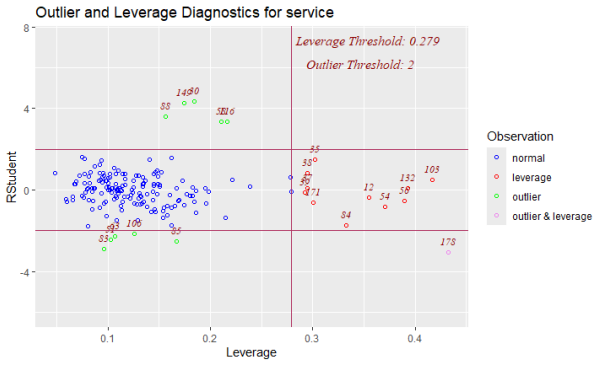

Figure 8: Potential residual plot iteration 2

Final Model and Conclusion

A final model was built using only the significant variables from the initial analysis. The `lm()` function was used to fit the model. For the variable selection, ols_step_backward_aic was tested.

The stepwise regression process led to the selection of 10 significant variables for the final model, based on the Akaike Information Criterion (AIC), Schwarz Bayesian Criterion (SBC), and Adjusted R-Squared (R^2_{adj}) values. Starting from a base model with no predictors, the inclusion of variables like `end_of_month`, `weekend`, `weather_main_Rain`, and various weekday indicators improved the model's fit, as indicated by the decrease in AIC and SBC values.

In the final model, we achieved an R^2 of 0.512 and an Adjusted R^2 of 0.483, suggesting a moderate fit of the model to the data. The model's performance is also supported by an RMSE of 31.863 and an MAE of 21.967, indicating a reasonable level of accuracy in predictions. It was also

noted the absence of the variable `service_day_trend` which is reasonable as its addition reduced the adjusted R^2.

The diagnostics of the final model shows similar graph interpretation as iteration 2 indicating the same generizability with less predictors; and thus a better final model.

References

[1] Christian Rudloff, Maximilian Leodolter, Dietmar Bauer, Roland Auer, Werner Brög, and Knud Kehnscherper. Influence of weather on transport demand: A case study from the vienna region. *ResearchGate*, 2015.

[2] A.P. Sandqvist and B. Siliverstovs. Is it good to be bad or bad to be good? assessing the aggregate impact of abnormal weather on consumer spending. *Empirical Economics*, 61:3059–3085, 2021.