Predicting Volume of Rideshare Trips

Time Series Analysis and Forecasting

Duo Zhou

Date:
Dec 7th, 2020

Motivation and Objective of the Project

Motivation

- Rideshare providers such as Uber and Lyft are an **important means of transportation in big cities** such as Chicago and can vary greatly by season.
- Given the comfort they offer to riders, their demand is **greatly influenced by weather conditions** such extreme heat, rainfall, wind, or cold temperatures

Objective of the Project

- The goal of this project is to **forecast the volume of rideshare trips.**
- **Weather data** will also be included with the purpose of understanding if and how it influences ridership
- The scope is limited to the city of **Chicago**
- The most straightforward **business application** of this project is to help optimize Rideshare Supply and Demand allocation

Agenda

- Motivation and Objective of the Project
- Data Sources and Data Preparation
- Main Characteristics of the Time Series
- Model Fitting
- Model Selection and Model Evaluation
- Conclusions and Next Steps

Data Sources and Data Preparation

Data Sources

Rideshare Data

- Obtained from the Chicago Data Portal (here)

Weather Data

- Obtained from National Centers for Environmental Information (here)

Data Preparation

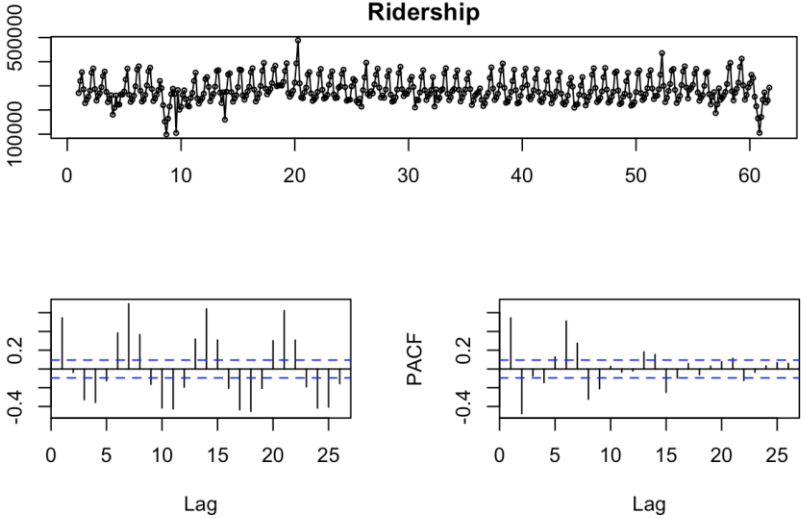
- Original Data :
  - 50 GB, Structured CSV File
  - 129 Million Rows, 21 Columns
- Processed Ridership Data
  - Ridership data aggregated and grouped by date
  - 426 Rows, 1 Column
  - Nov 2018 to Oct 2019 (Train)
  - Nov 2019 to Dec 2019 (Test/Forecast Horizon)
- Hourly Weather datasets
- Processed Weather Data
  - Data are aggregated and grouped by date
  - Mean Temperature, Precipitation and Wind Speed
  - 426 Rows and 3 Columns
  - Nov 2018 to Oct 2019 (Train)
  - Nov 2019 to Dec 2019 (Test /Forecast Horizon)

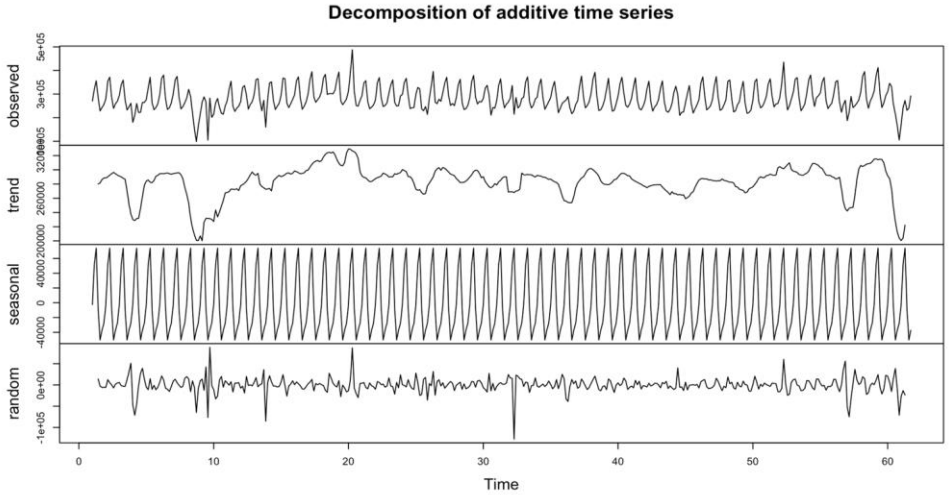
Agenda

- Motivation and Objective of the Project
- Data Sources and Data Preparation
- **Main Characteristics of the Time Series**
- Model Fitting
- Model Selection and Model Evaluation
- Conclusions and Next Steps

Main Characteristics of the Time Series (1/2)

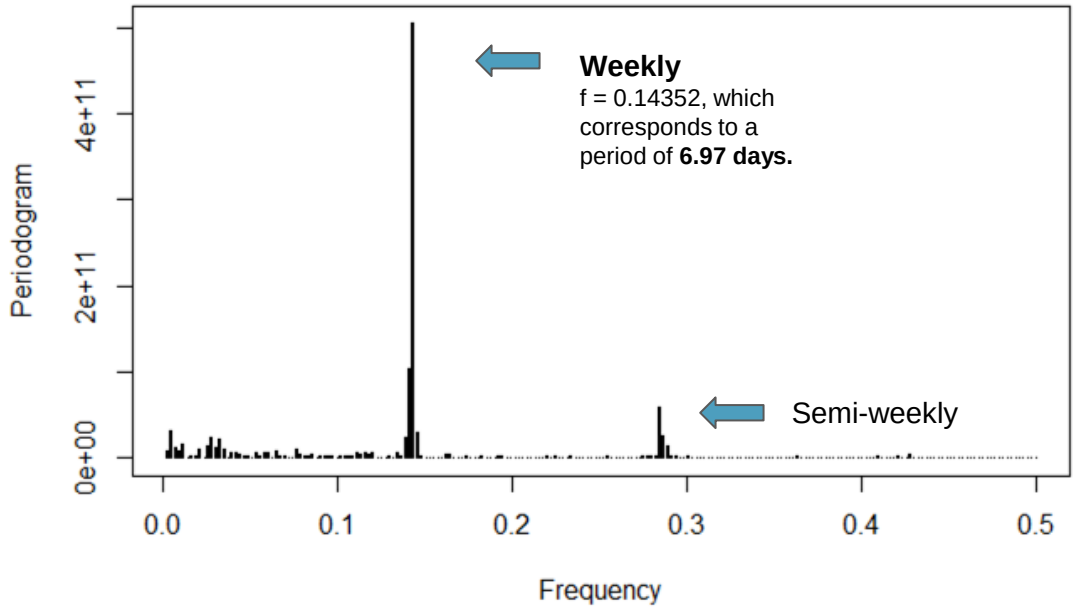
- Clear weekly seasonal pattern
- No need for Box-Cox
- No clear trend
- ADF Test - Data points are non-stationary





Main Characteristics of the Time Series (2/2)

- Our spectral analysis shows that, by far, the most significant seasonal component is weekly



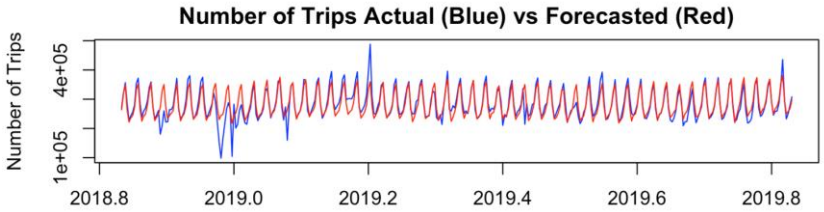
Agenda

- Motivation and Objective of the Project
- Data Sources and Data Preparation
- Main Characteristics of the Time Series
- **Model Fitting**
  - Linear Regression
  - Holt-Winters
  - sARIMA
  - Regression w/ ARIMA errors
  - Vector AutoRegressive Model (VAR)
- Model Selection and Evaluation
- Conclusions and Next Steps

Model Fitting - Linear Regression

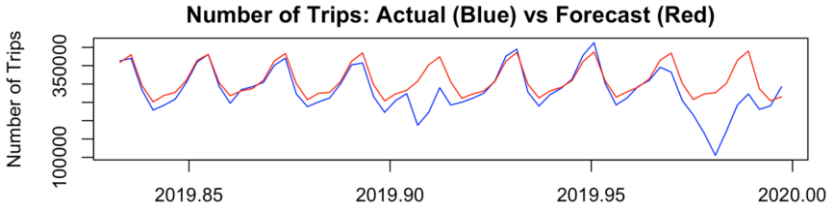
Modelling Phase Outline

- Asides from Time, **other predictors** used were: Temperature, Precipitation, and Wind Speed
- Some **data transformations** were required:
  - Convert Precipitation to binary flag
  - Create Day of Week variable
  - Create Day of Month variable
  - Create End of Month binary flag
- After optimizing AIC, Wind Speed and Day of Month were removed from predictors, yielding an **R2 of 72%**



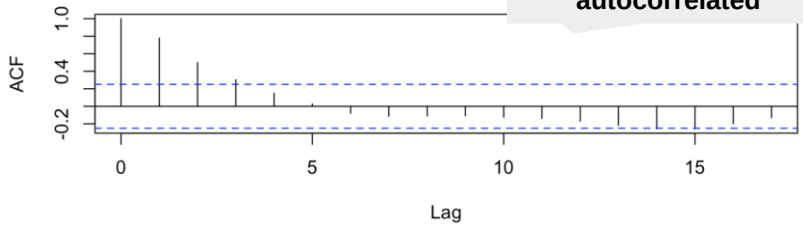
Results on Test Data

RMSE: 49,771
MAE: 31,171



ACF of Residuals:

Residuals are **autocorrelated**

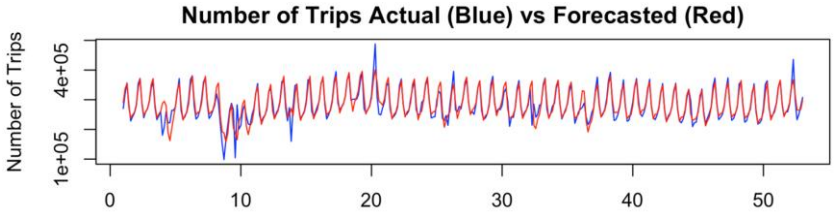


Ljung-Box Test:
Residuals are NOT independent (p-value = 1.135e-09)

Model Fitting - Holt-Winters

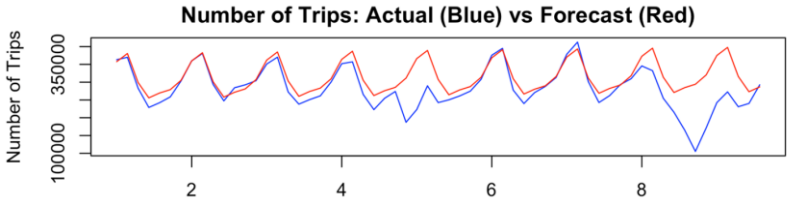
Modelling Phase Outline

- **Weekly seasonality** was modeled by setting the TS object with Frequency = 7
- Seasonality was also modeled as **Additive** after observing that Multiplicative yielded higher AICc
- The resulting **smoothing parameters** are:
  - Alpha (level) = 0.452
  - Beta (trend) = 0.003
  - Gamma (seasonality) = 1e-4



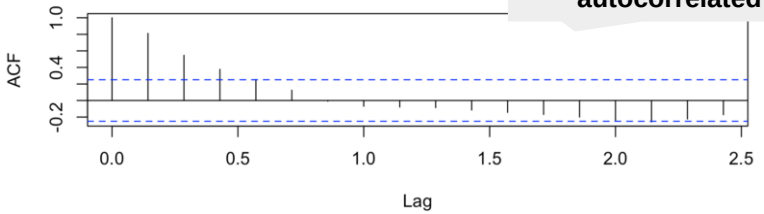
Results on Test Data

RMSE: 55,765
MAE: 35,141



ACF of Residuals:

Residuals are **autocorrelated**

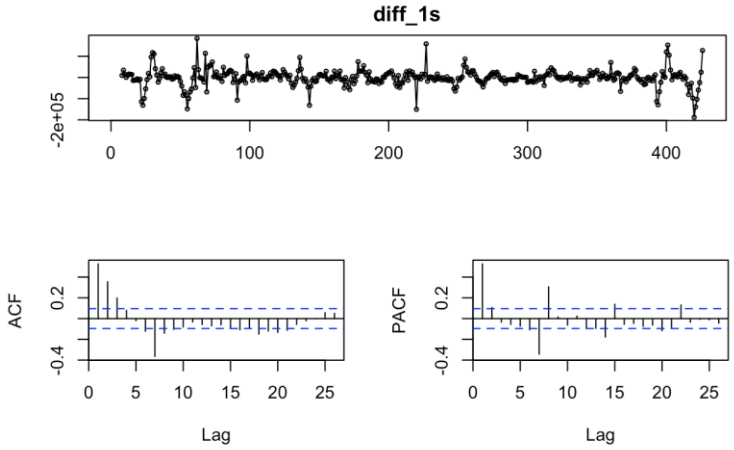


Ljung-Box Test:
Residuals are NOT independent (p-value = 2.31e-10)

Model Fitting - sARIMA

Modelling Phase Outline

- Order one of seasonal differencing



- Model Estimation - “Maximum Likelihood Estimation”
- sARIMA(1,0,0)(0,1,1)[7]

  sARIMA(2,0,0)(0,1,1)[7]

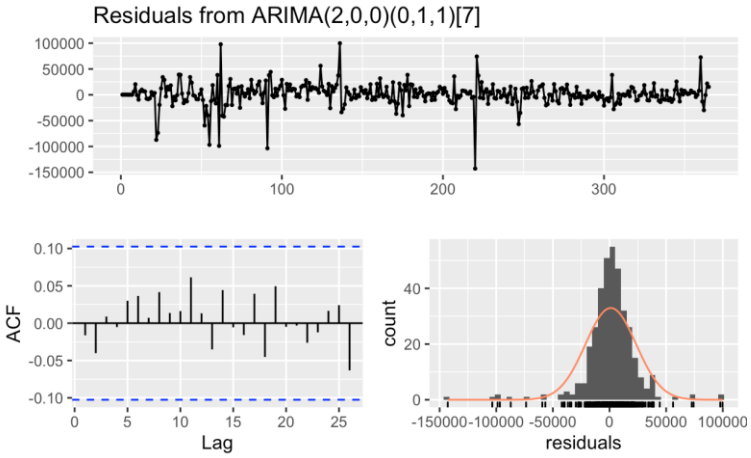
  ARIMA(2,0,1) - auto.arima

	AICc	BIC
Model 1	8246.2	8257.8
Model 2	8226.1	8241.5
Model 3	8746.8	8766.1

Model Fitting - sARIMA

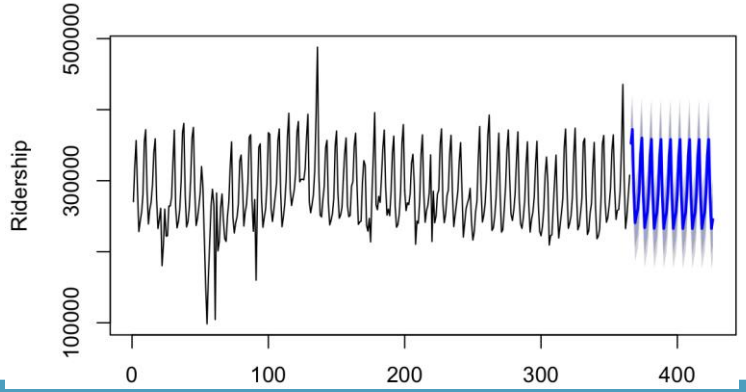
Model Diagnostic

- Model 1: Residuals of are not fully white noise
- Model 2: Residuals of are white noise
- Model 3: Residuals of are not fully white noise



Forecast

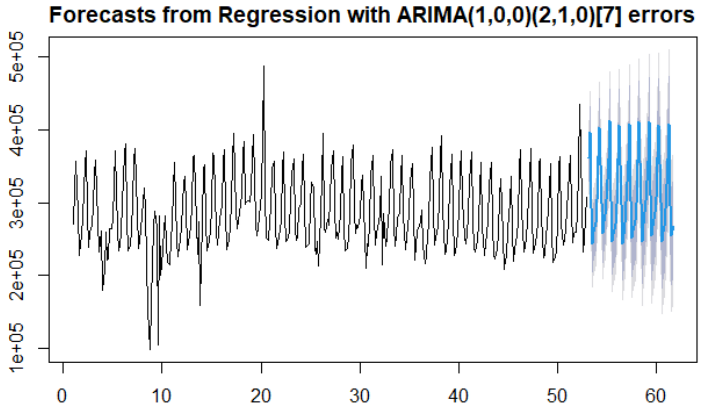
- **Simple Forecasting Methods:**
  - Average Method
  - RMSE: 51407
  - MAE: 40426
- **Forecast using Model 2**
  - RMSE: 42979.81
  - MAE: 30065.31



Model Fitting - Regression w/ auto.arima errors

Modelling Phase Outline

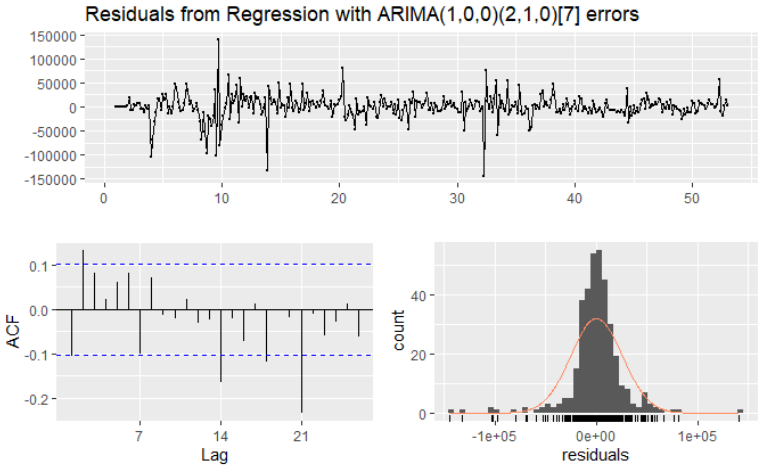
- Fitted using auto.arima
- Resulted in ARIMA(1,0,0)(2,1,0)[7] for errors
- Residuals are not considered time-independent



Results on Test Data

RMSE: 50,484
MAE: 30,115

ACF of Residuals:



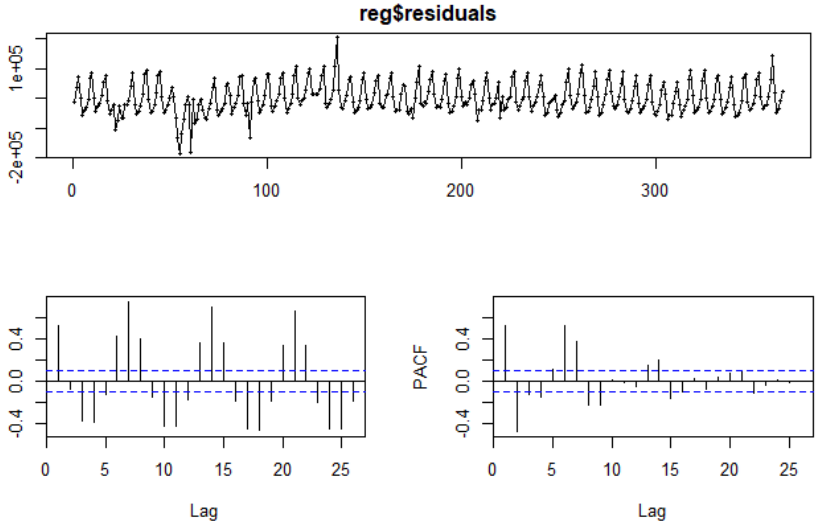
Ljung-Box Test:

Residuals are NOT independent (p-value = 1.559e-05)

Model Fitting - Regression w/ ARIMA errors

Modelling Phase Outline

- Fitted linear regression to analyze errors:



- As with previous sARIMA model, we identified an AR model with seasonal component most likely at 7
- We will fit the sARIMA for the errors given the best previous sARIMA model.

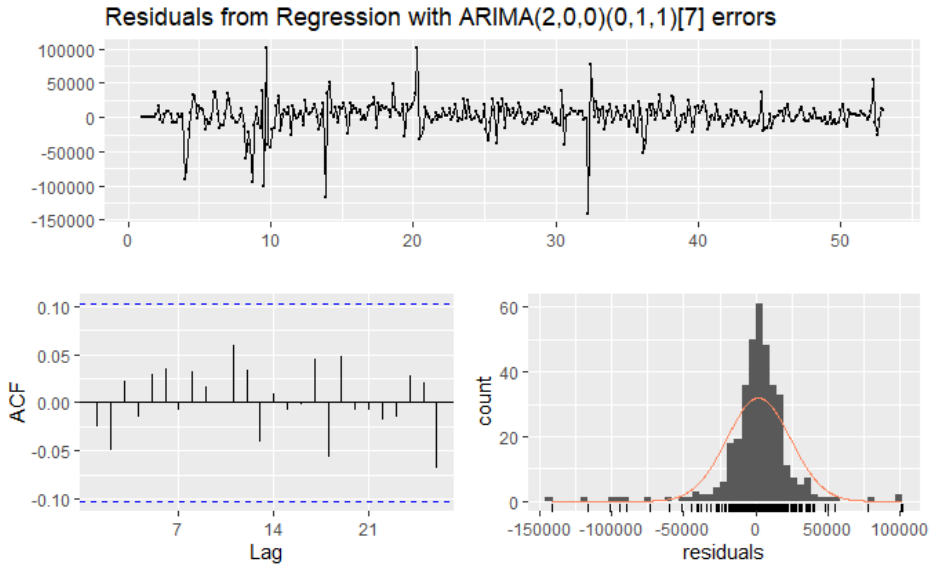
Model Fitting - Regression w/ ARIMA errors

Results on Test Data

RMSE: 42,678
MAE: 27,021

Ljung-Box Test:
Residuals ARE independent (p-value = 0.7368)

ACF of Residuals:



Model Fitting - VECTOR AUTOREGRESSIVE MODEL (VAR)

Modelling Phase Outline

- Use VARselect to choose P orders, Season=7

```
AIC(n)  HQ(n)  SC(n) FPE(n)
     3       1      1      3
```

- Based on the selection above we run two models: VAR(1) and VAR(3). Season =7

```
VAR1<-VAR(mv.ts, p = 1, season = 7, type='both')
VAR3<-VAR(mv.ts, p = 3, season = 7, type='both')
```

- Then we run the serial.test on both models and found that there is no autocorrelation in the residuals of VAR(3) but there are auto correlations in the residuals of VAR(1).

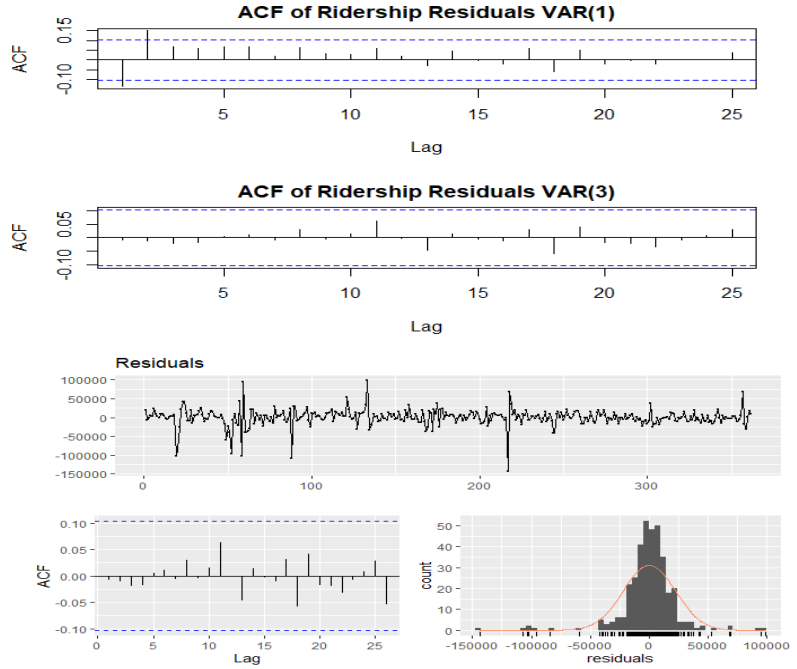
```
        Portmanteau Test (asymptotic)

data:  Residuals of VAR object VAR1
Chi-squared = 175.26, df = 144, p-value = 0.03905


        Portmanteau Test (asymptotic)

data:  Residuals of VAR object VAR3
Chi-squared = 104.81, df = 112, p-value = 0.6724
```

ACF of Residuals:



VAR(3) Residuals

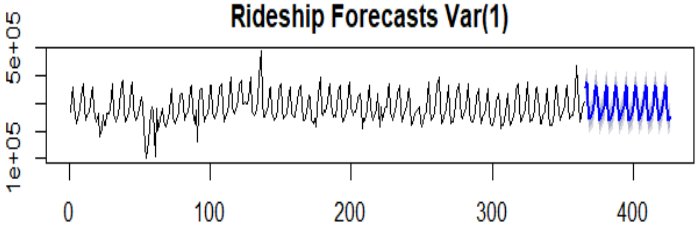
Model Fitting - VECTOR AUTOREGRESSIVE MODEL (VAR)

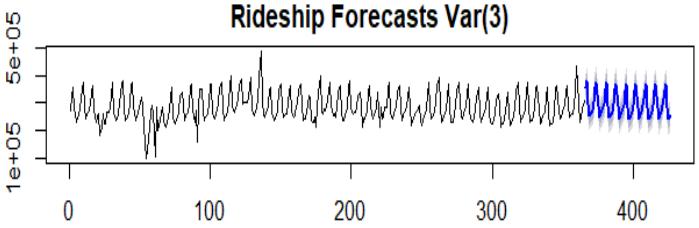
Results on Test Data

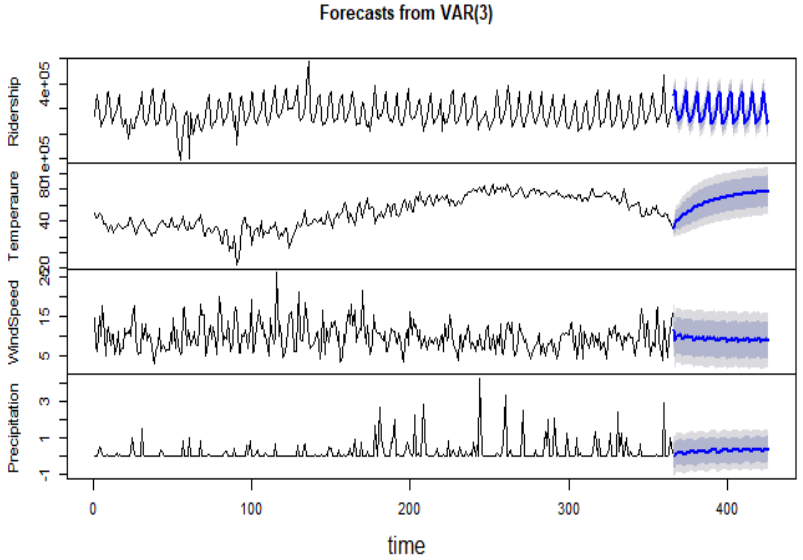
Ridership Eval Measures

RMSE:
VAR(1): 44,555
VAR(3): 43,789

MAE:
VAR(1): 28,403
VAR(3): 27,891



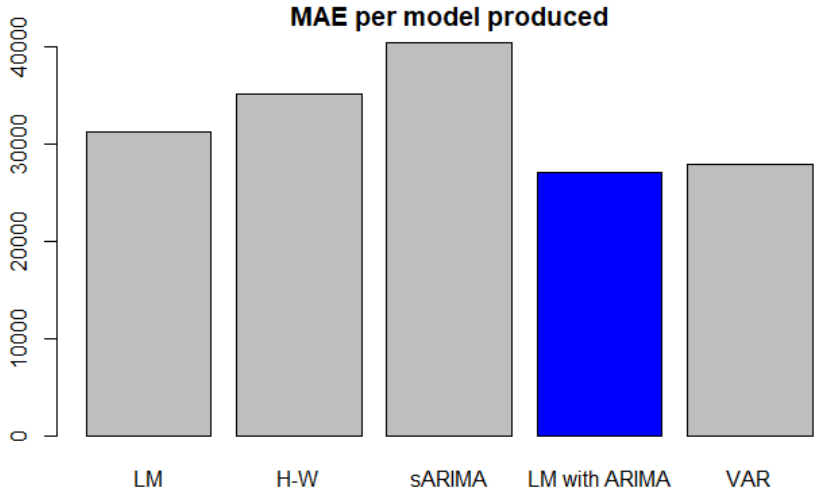




Agenda

- Motivation and Objective of the Project
- Data Sources and Data Preparation
- Main Characteristics of the Time Series
- Model Fitting
- **Model Selection and Model Evaluation**
- Conclusions and Next Steps

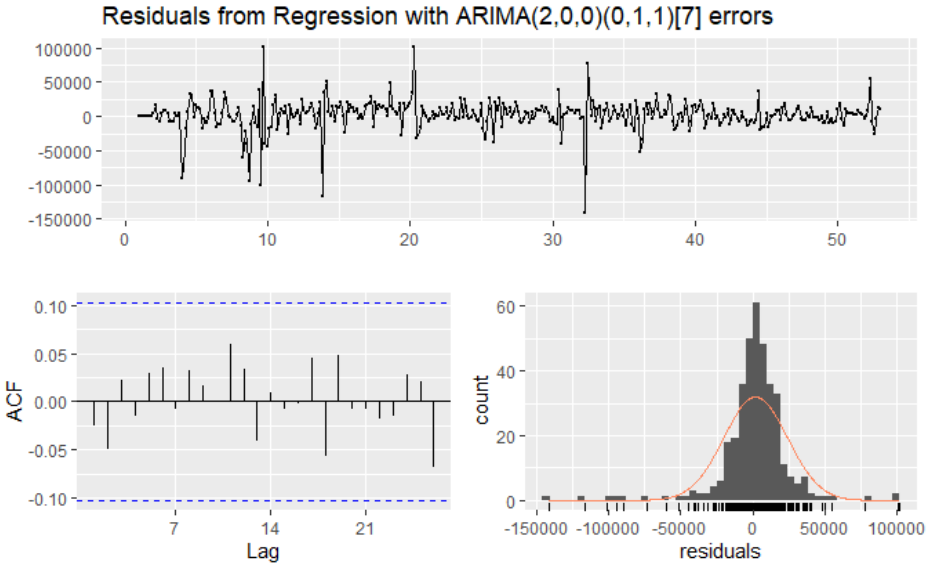
Model Selection



- Not surprisingly, LM with ARIMA and VAR were the best models, coming very close to each other, because they both considered the weather variables that we would expect to influence uber ridership.
- However, a simple LM also did a great job at predicting even though its errors were not white noise at all!

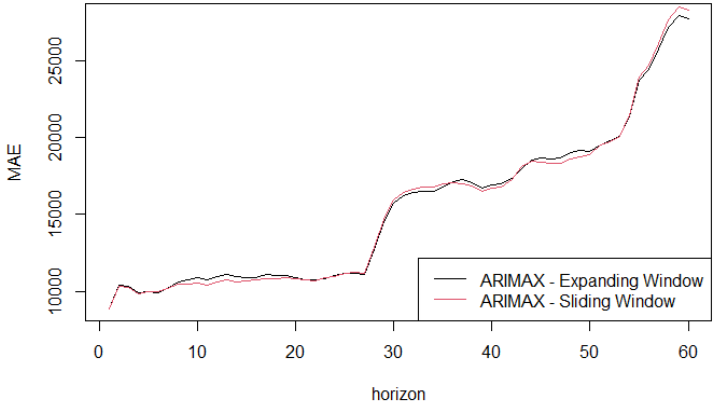
Chosen Model Evaluation

ACF of Residuals:



Cross-validation

Errors are stable until holiday periods are introduced, the two sharp rises of errors were caused by Thanksgiving and Christmas



Ljung-Box Test:

Residuals ARE independent (p-value = 0.7368)

Agenda

- Motivation and Objective of the Project
- Data Sources and Data Preparation
- Main Characteristics of the Time Series
- Model Fitting
- Model Selection and Model Evaluation
- **Conclusions and Next Steps**

Conclusions and Next Steps

Conclusions

In conclusion, multivariate models performed better than univariate models, signaling the **importance of weather related variables** in the prediction of rideshare trips

Next Steps

As next steps, the VAR model could be expanded to VARMA and sVARMA models to better address stationarity and cross-correlations

- A Coefficient Restriction Matrix could be applied to eliminate the influence of rideshare trips on weather features
- This means, only weather would influence ridership but not the other way around
- Interventions should be introduced for holiday seasons.

Thank you!

Project work distribution

- Presentation Design + Objective of the Project - Lucia
- Data Sources and Data Preparation - Amily / Duo
- Main Characteristics of the Time Series - Amily / Luis / Duo
- Linear Reg. + Holt Winters - Lucia
- sARIMA - Amily
- Regression with ARIMA errors - Luis
- VARs - Duo
- Model Selection and Model Evaluation - Luis
- Conclusions and Next Steps - Lucia / Duo