DATA ANALYSIS AND FORECASTING FOR HOUSE PRICING DATA

Elif Doğan Dar

December, 2021

DATA DESCRIPTION

<u>unit_master</u> is a data frame with 1100 observations and 5 variables.

A - UnitNumber (int): Unique number given to the rental.

B - UnitType (factor with 3 levels): Studio Apartment (STD) , 1 Bedroom Apartment(1BR) and 2 Bedroom Apartment(2BR)

C - UnitPlan (factor with 11 levels): UnitType is further divided into subcategories. There are 100 observations per UnitPlan.

STD: STD-L, STD-M&A, STD-M&B, STD-S&A, STD-S&B, STD-S&C

1BR: 1BR-L&A, 1BR-L&B, 1BR-S&A, 1BR-S&B

2BR: 2BR-S

D - Sqft(int): Area of the rental. It has 6 different values

450 for STD-S&A, STD-S&B, STD-S&C

650 for STD-M&A STD-M&B

850 for STD-L

750 for 1BR-S&A 1BR-S&B

1050 for 1BR-L&A 1BR-L&B

1250 for 2BR-S

E - Floor(int): Floor of the rental taking values between 1 to 10, it does not affect the pricing directly.

DATA DESCRIPTION

rental_master is a data frame with 7236 observations and 5 variables.

LeaseNo (int) : Unique number given to lease of the rental.

CustomerNo (int) : Unique number given to the customer.

UnitNumber (int) : It connects rental_master to unit_master.

StartDate (Date) : Start date of the lease. Values are between 2012-03-08 and 2021-03-06.

EndDate (Date) : End date of the lease. Values are between 2014-12-01 and 2021-03-06. If the lease didn't end till 2021-03-06, then this value is NA.

unit_rent_master is a data frame with 836 observations of 4 variables. It gives rise to 11 time series with a window of 76 months.

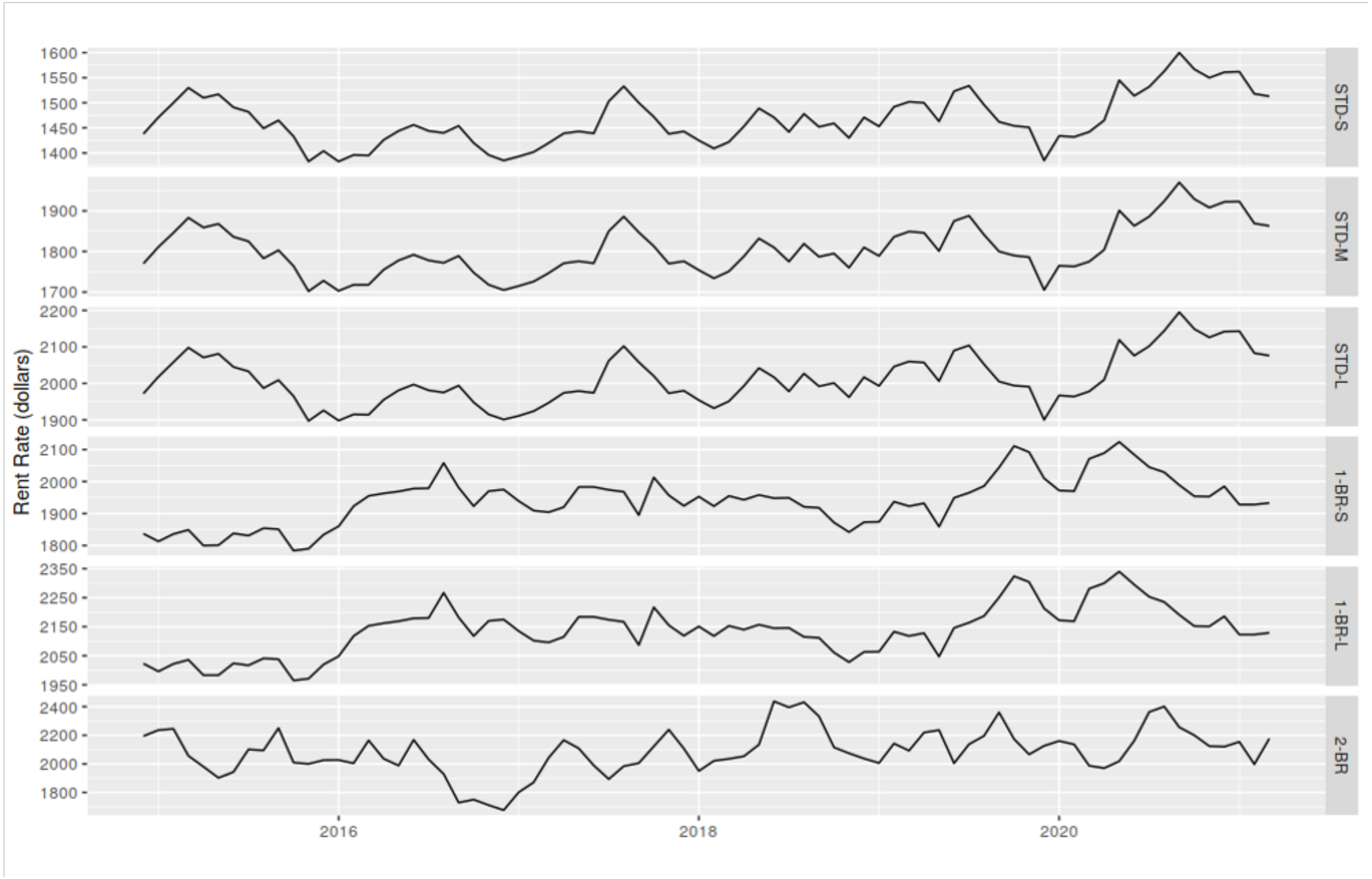
UnitPlan (factor with 11 levels): It connects unit_rent_master to unit_master.

StartDate (Date) : Start date of the particular month of the pricing (between 2014-12-01 and 2021-03-01).

EndDate (Date) : End date of the particular month of the pricing (between 2014-12-31 and 2021-03-31).

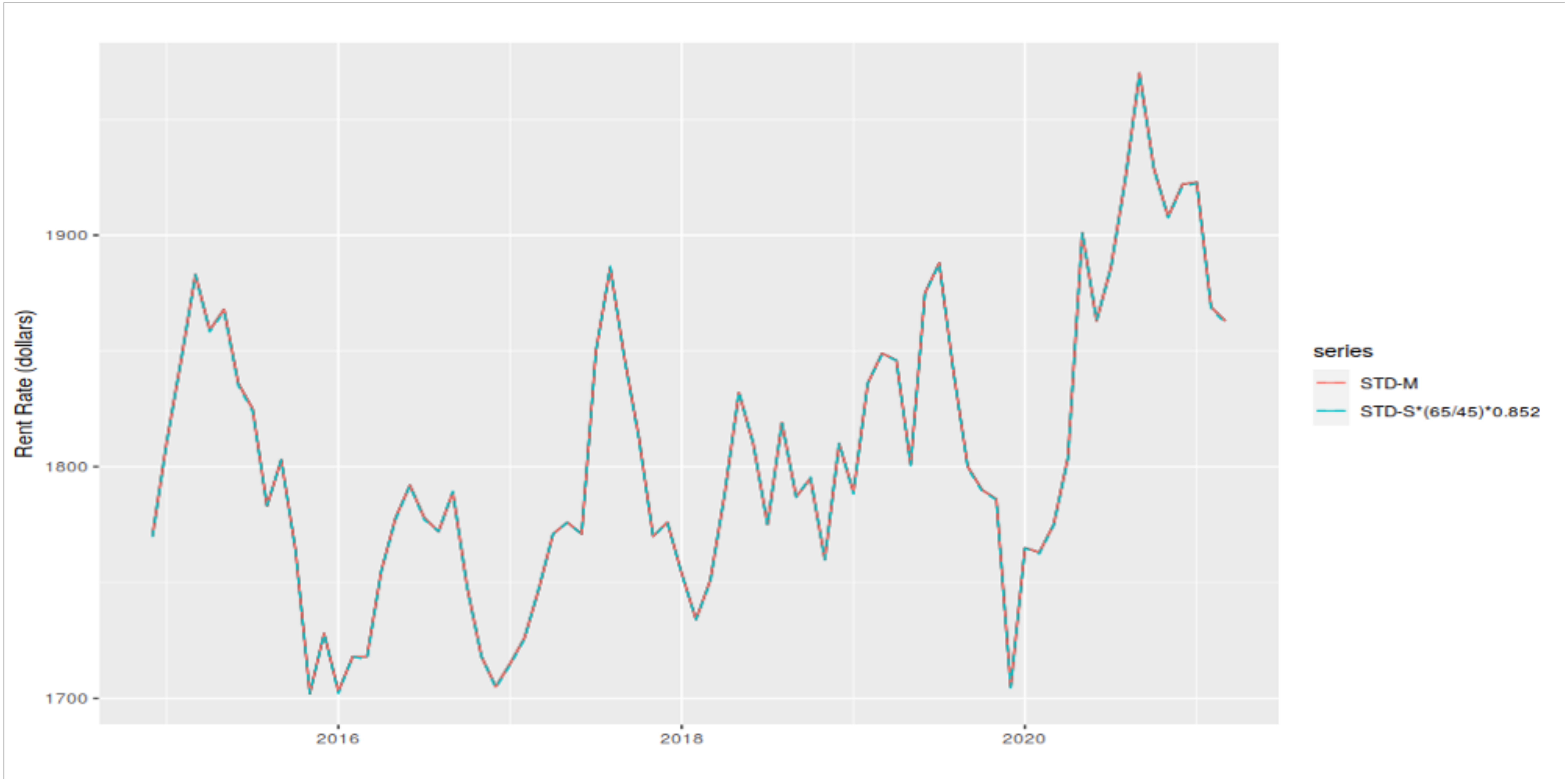
RentRate (int) : Rent rate in dollars for the particular month.

TIME SERIES

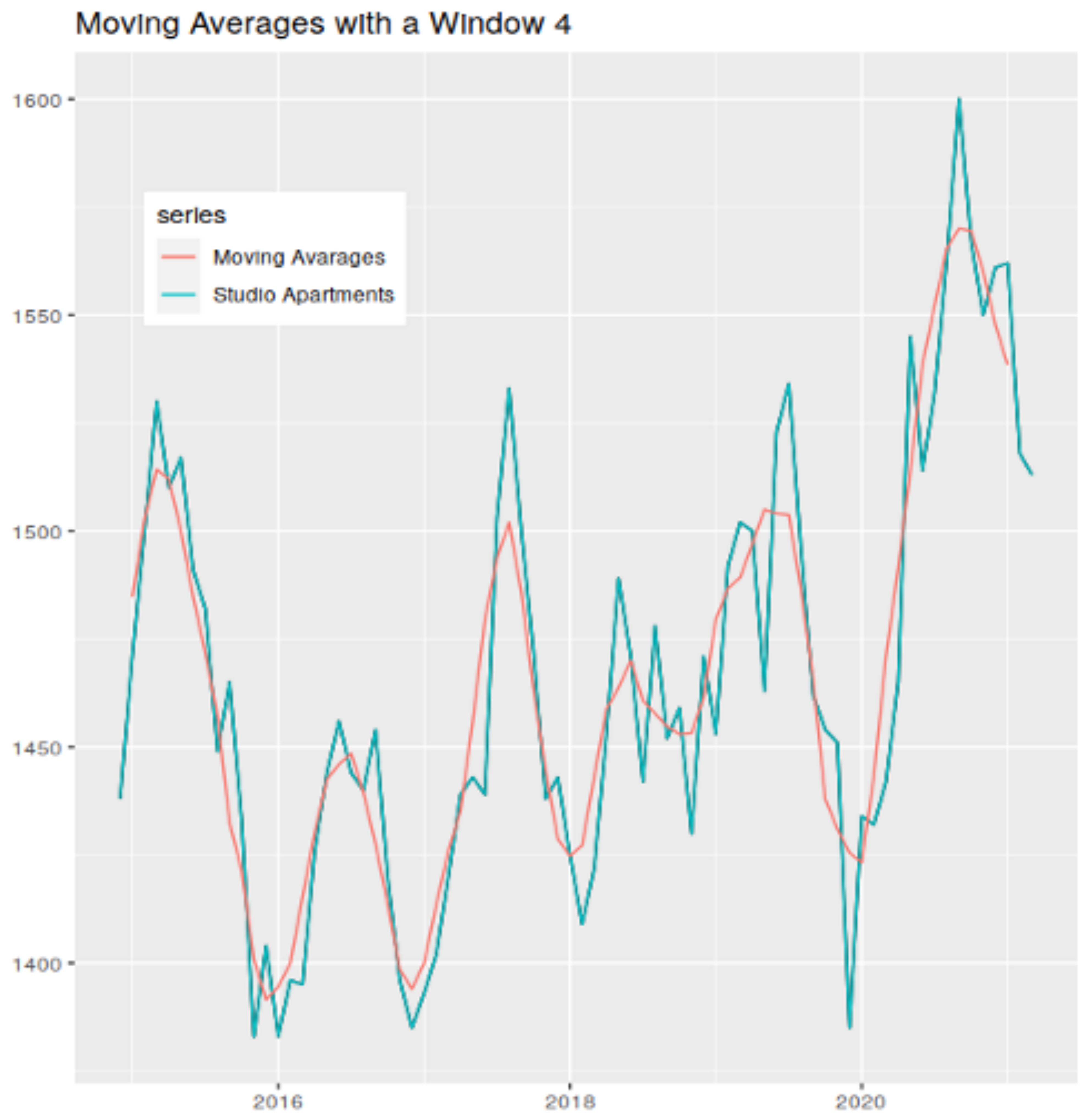


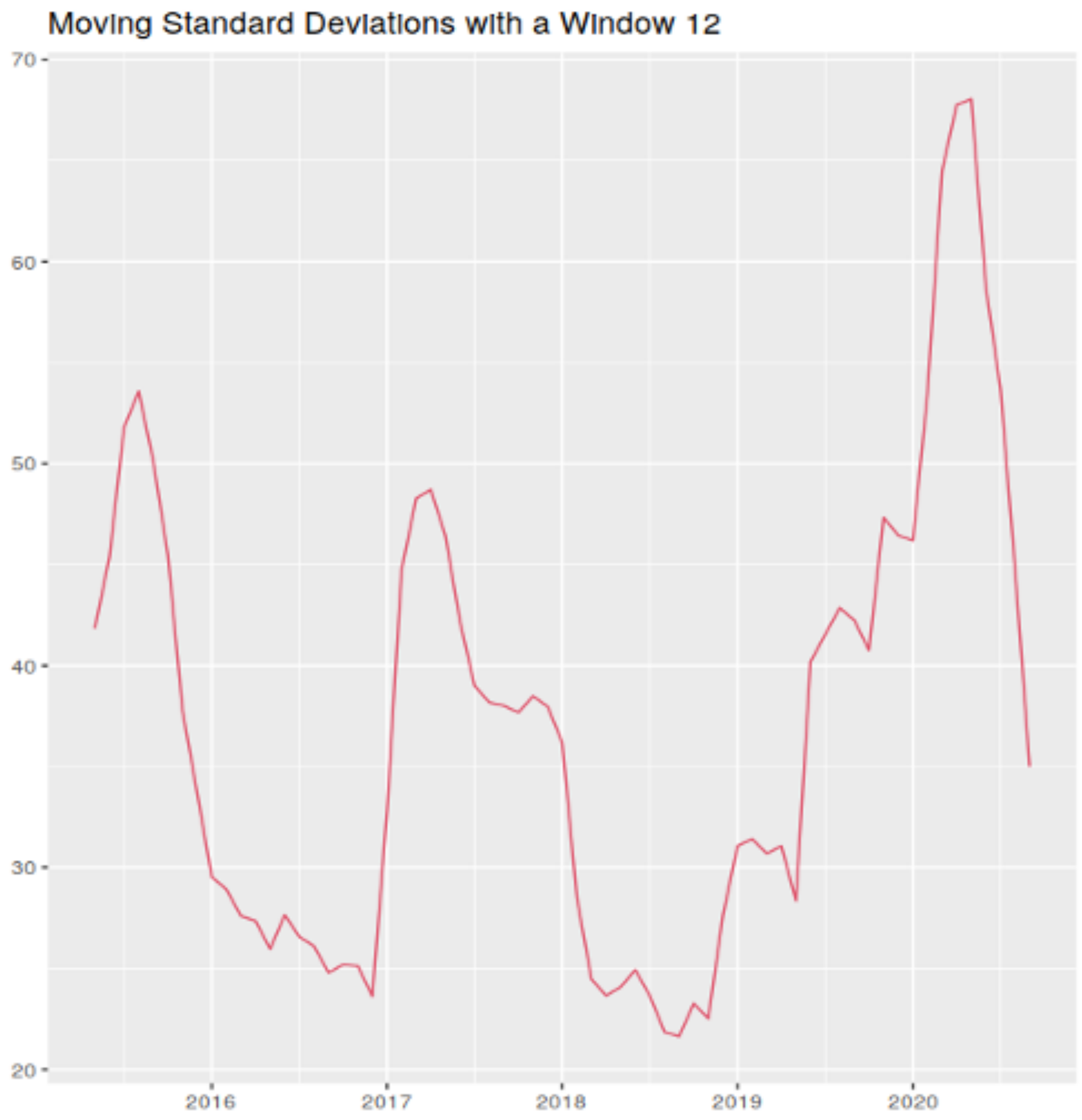
Rent Rate (dollars)
1600
1550
1500
1450
1400
STD-S
1900
1800
1700
STD-M
2200
2100
2000
1900
STD-L
2100
2000
1900
1800
1-BR-S
2350
2250
2150
2050
1950
1-BR-L
2400
2200
2000
1800
2-BR
2016
2018
2020

RELATION BETWEEN STD-S AND STD-M



MOVING AVERAGES AND MOVING STANDARD DEVIATIONS





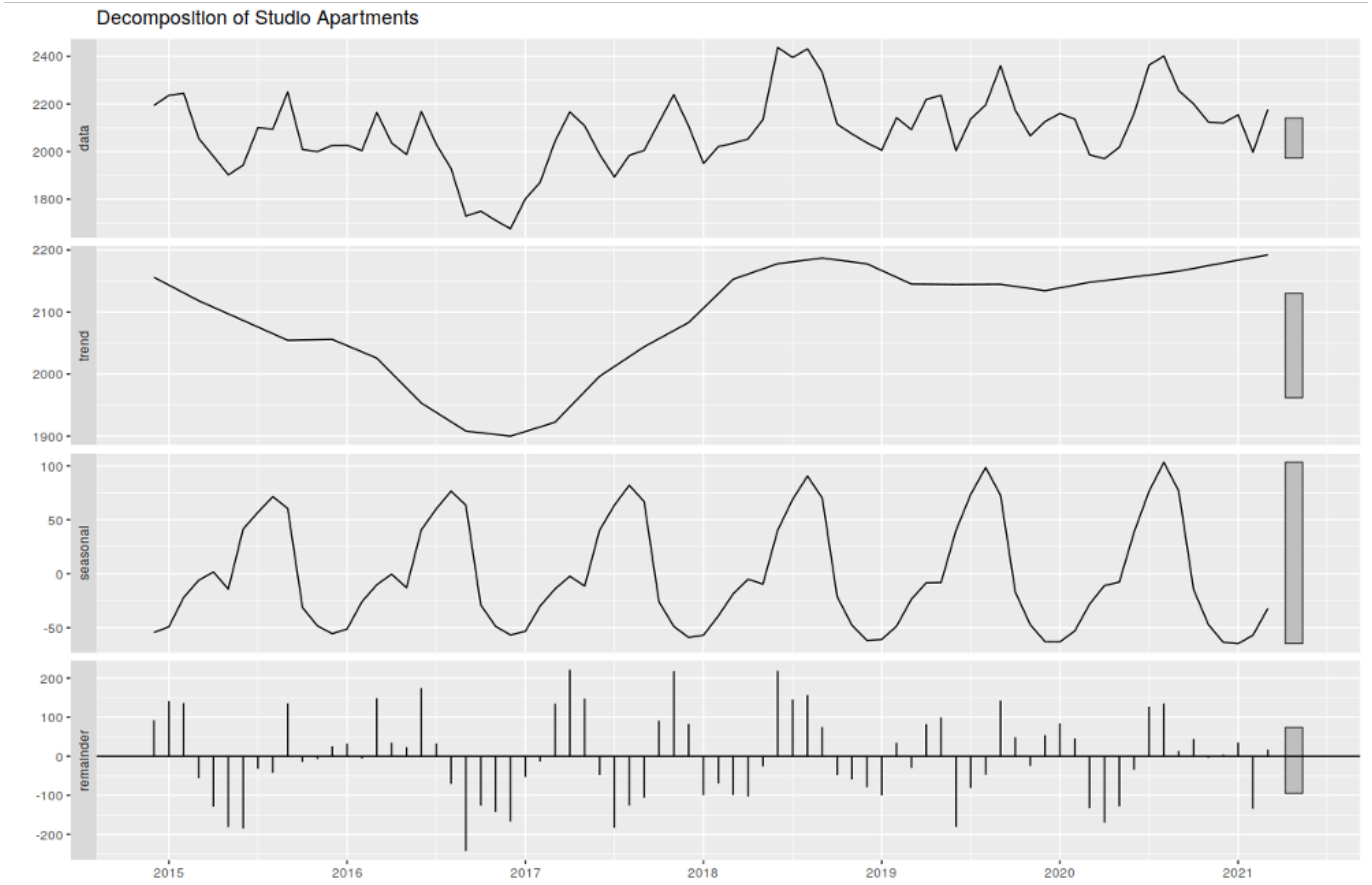
DECOMPOSITION OF STD

Data

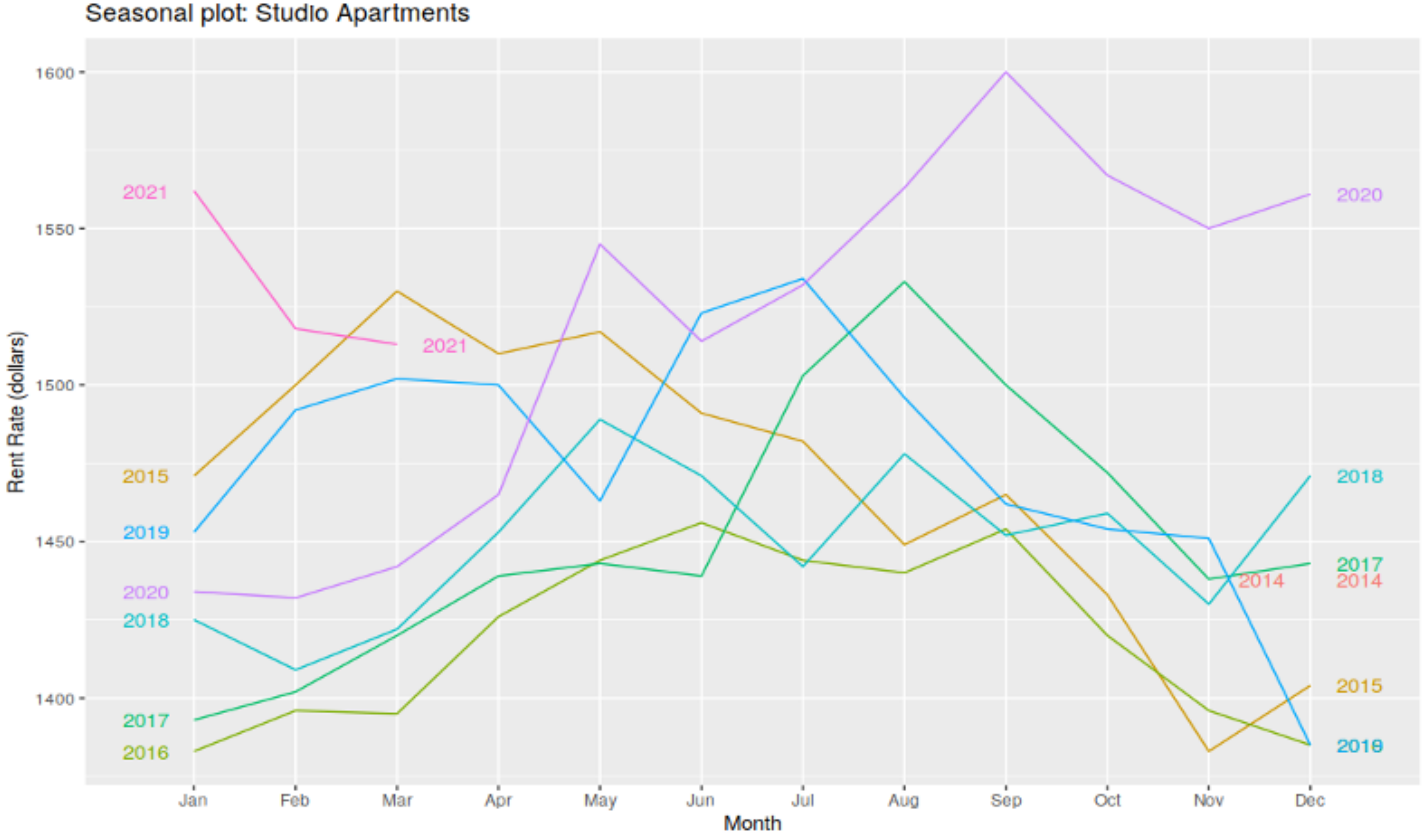
Trend

Seasonal

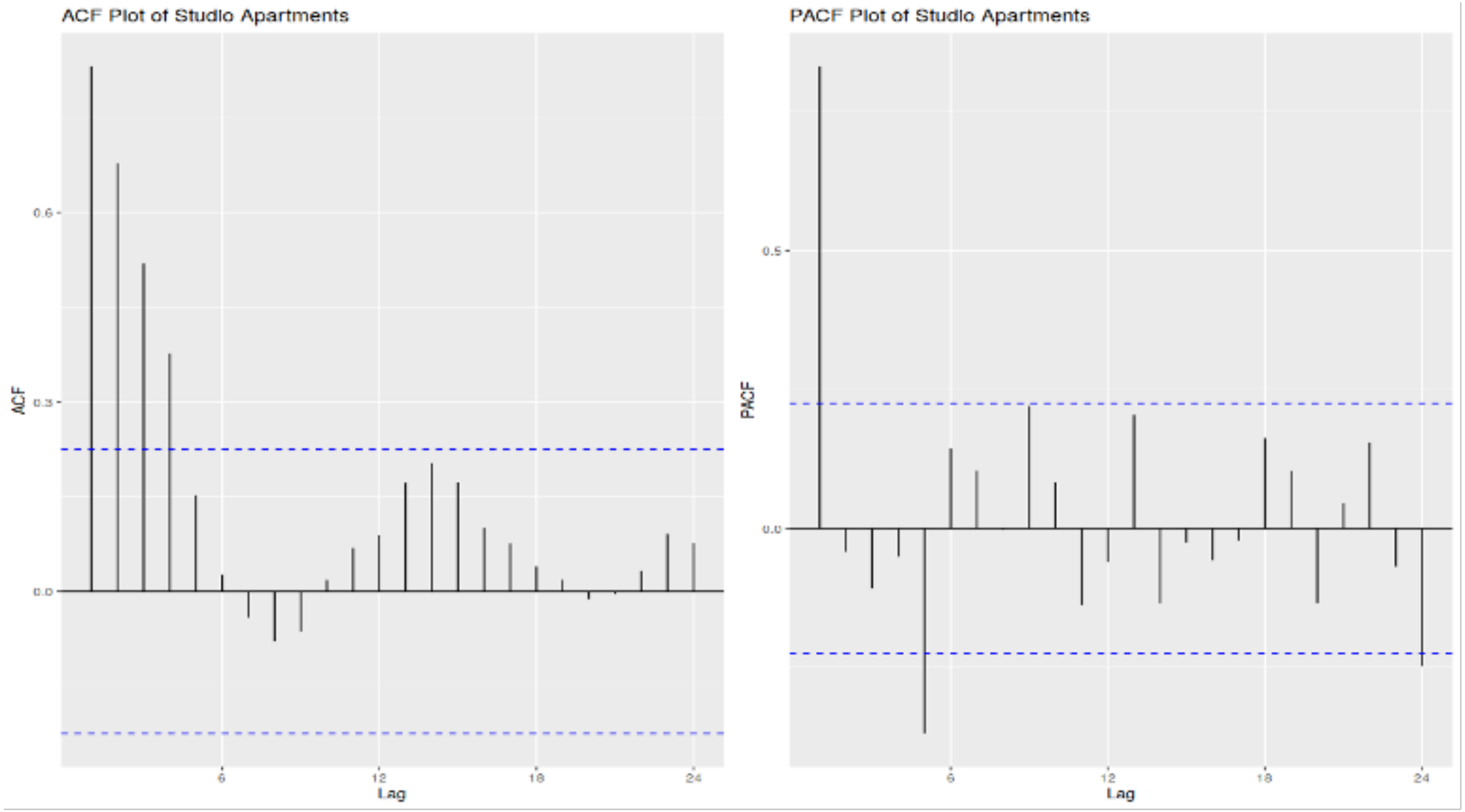
Remainder



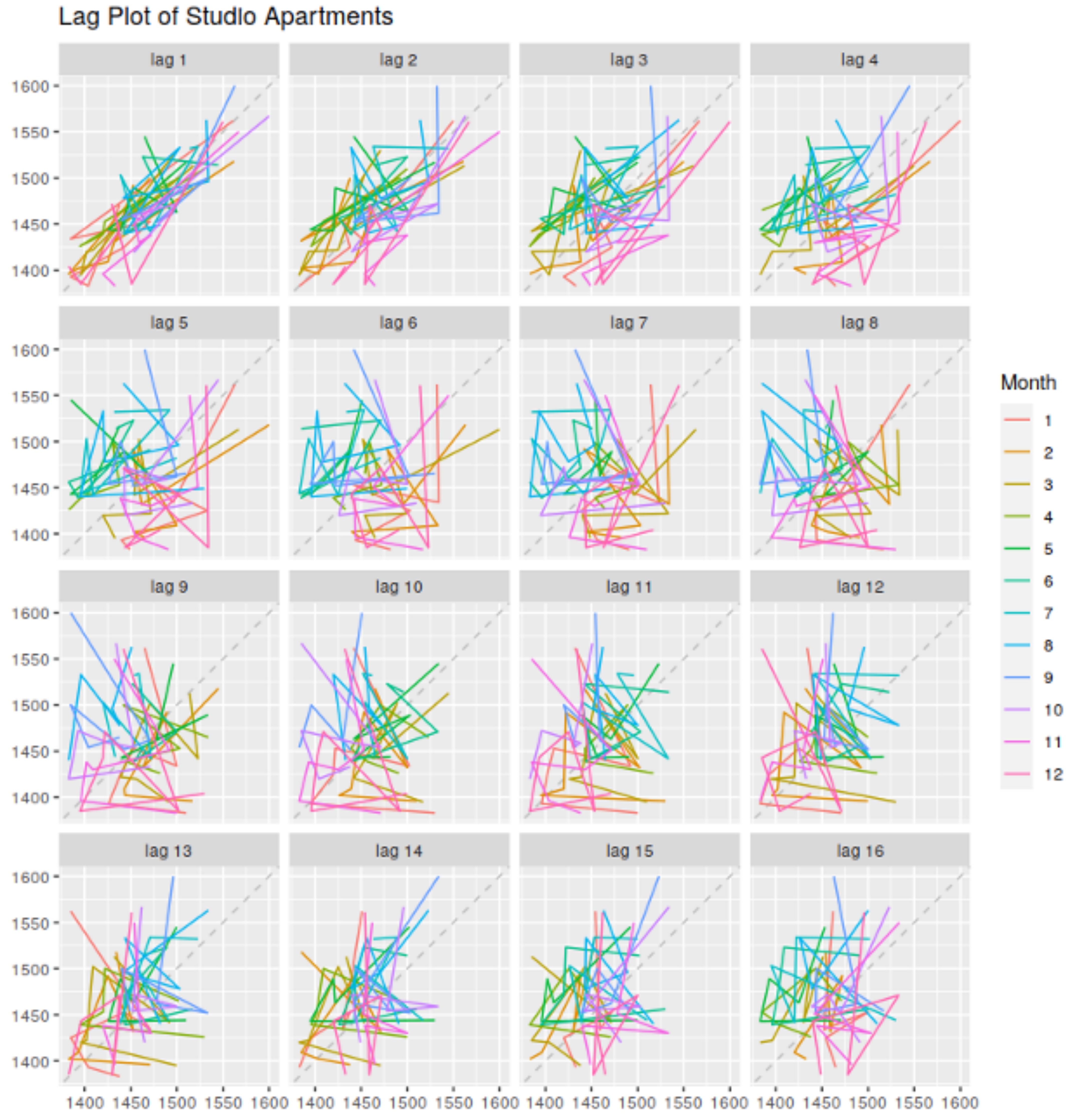
SEASONAL PLOT



AUTOCORRELATION AND PARTIAL AUTOCORRELATION PLOTS



LAG PLOT



Lag Plot of Studio Apartments
lag 1
lag 2
lag 3
lag 4
lag 5
lag 6
lag 7
lag 8
lag 9
lag 10
lag 11
lag 12
lag 13
lag 14
lag 15
lag 16
1400
1450
1500
1550
1600
Month
1
2
3
4
5
6
7
8
9
10
11
12

STATIONARITY

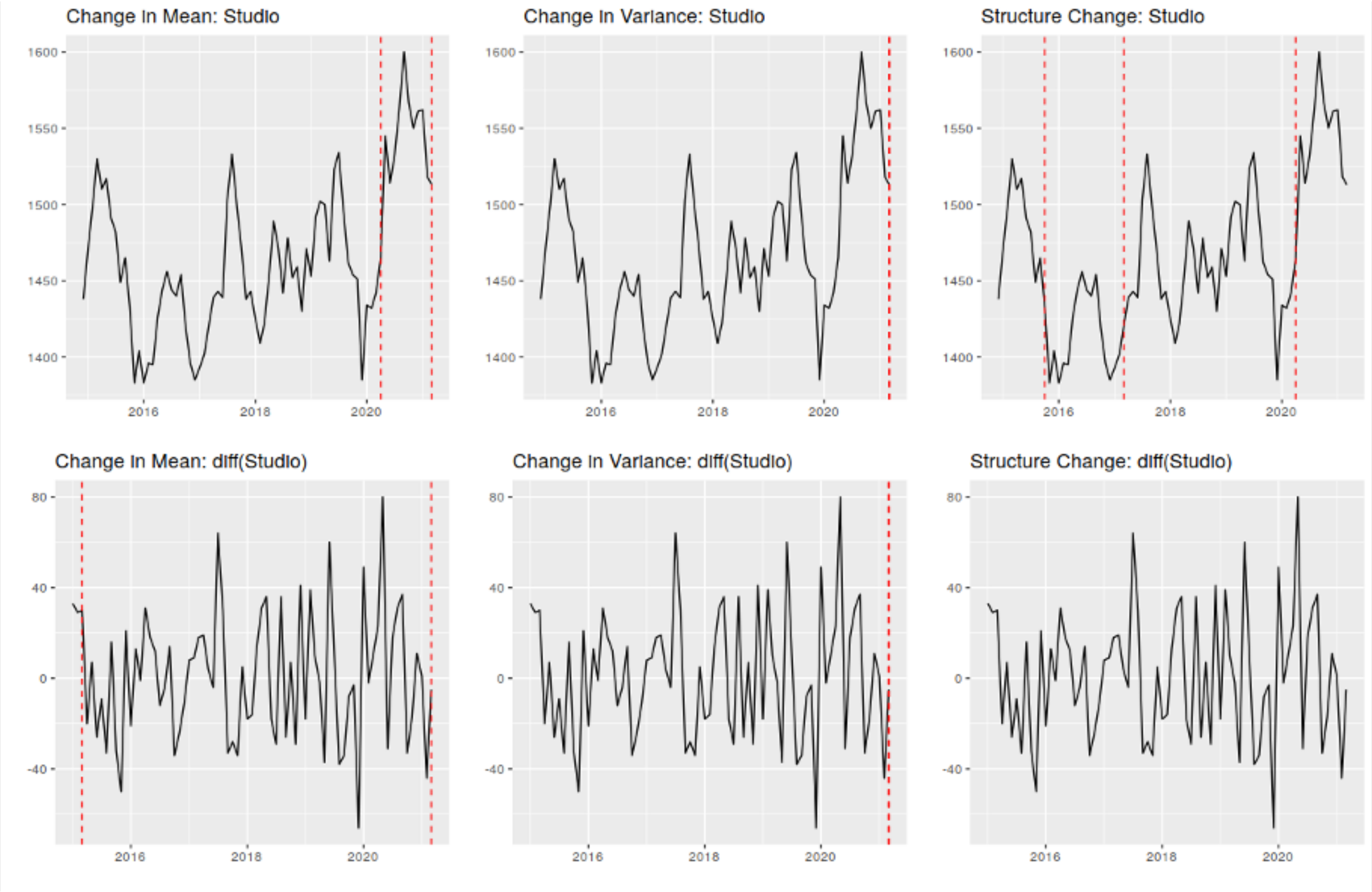
A time series is stationary if

1) It has constant mean over time.
2) It has constant standard deviation over time.
3) There is no autocorrelation.
4) There is no seasonality.

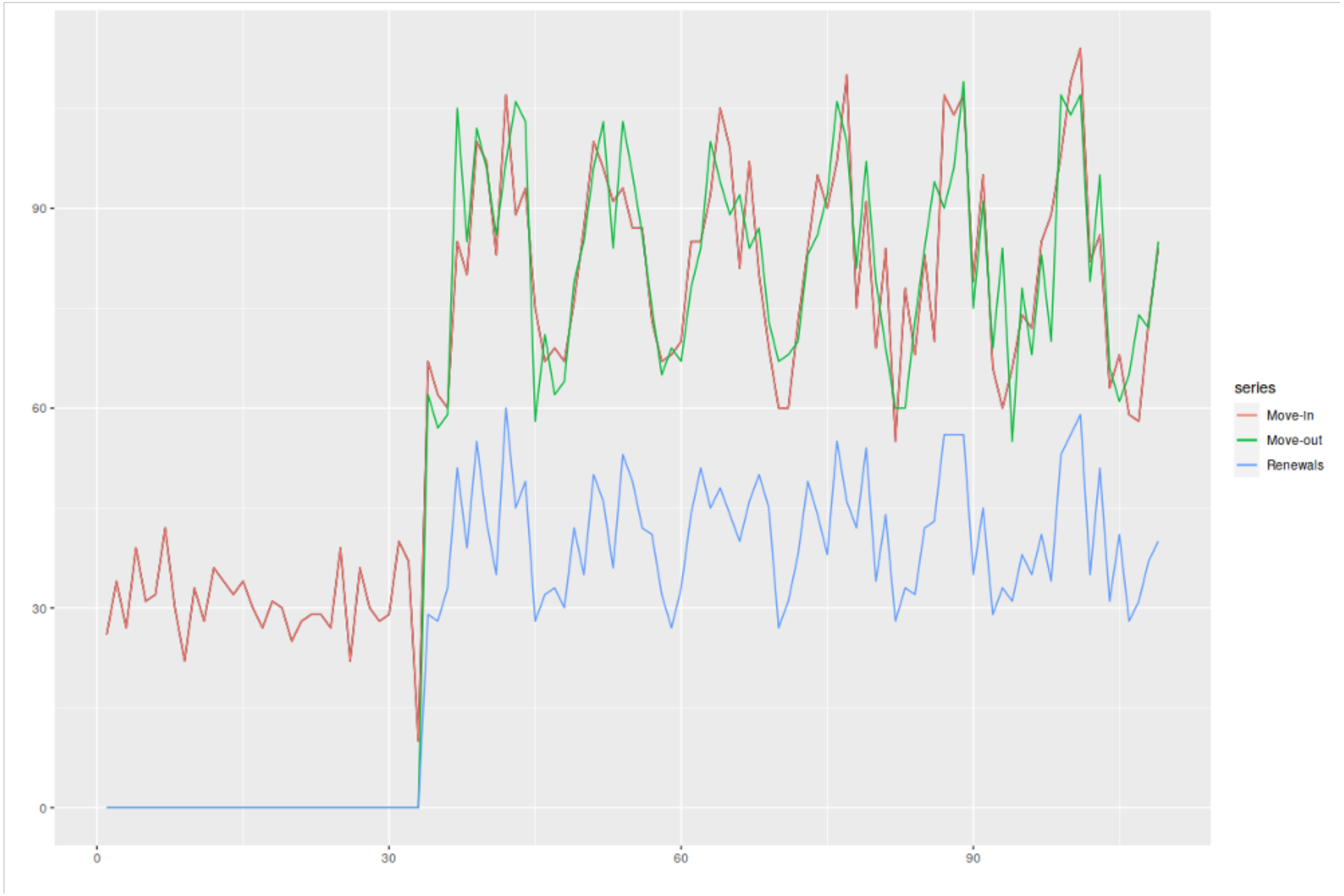
Stationarity Tests:

1) Augmented Dickey–Fuller Test (ADF). Stationary if the p-value is less than 0.05. (Stationary)
2) Phillips–Perron Unit Root Test . Stationary if p-value is less than 0.05. (Non-stationary)
3) Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. Stationary if p-value is greater than 0.05. (Stationary)

STATIONARITY



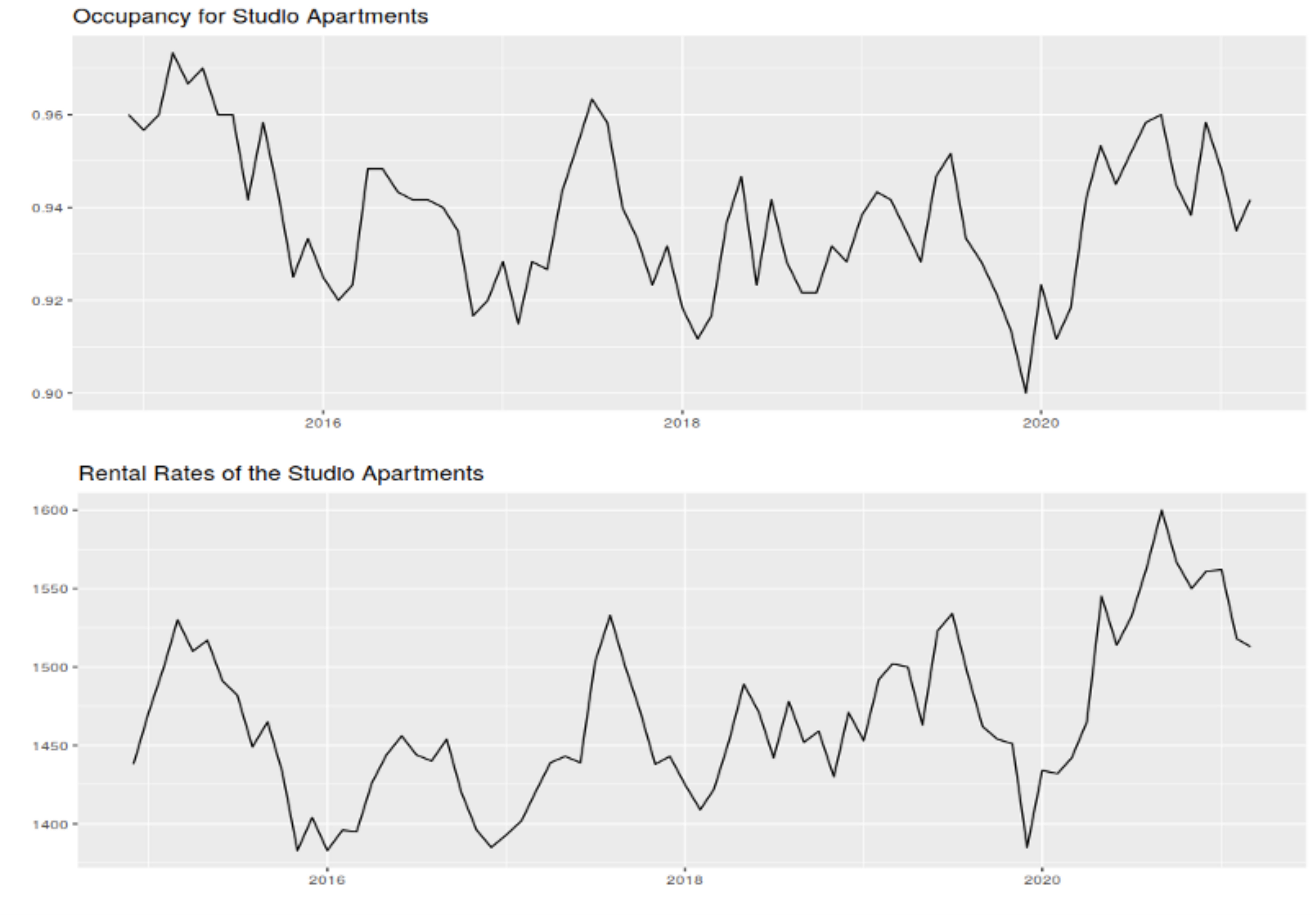
MOVE-IN, MOVE-OUT AND RENEWALS



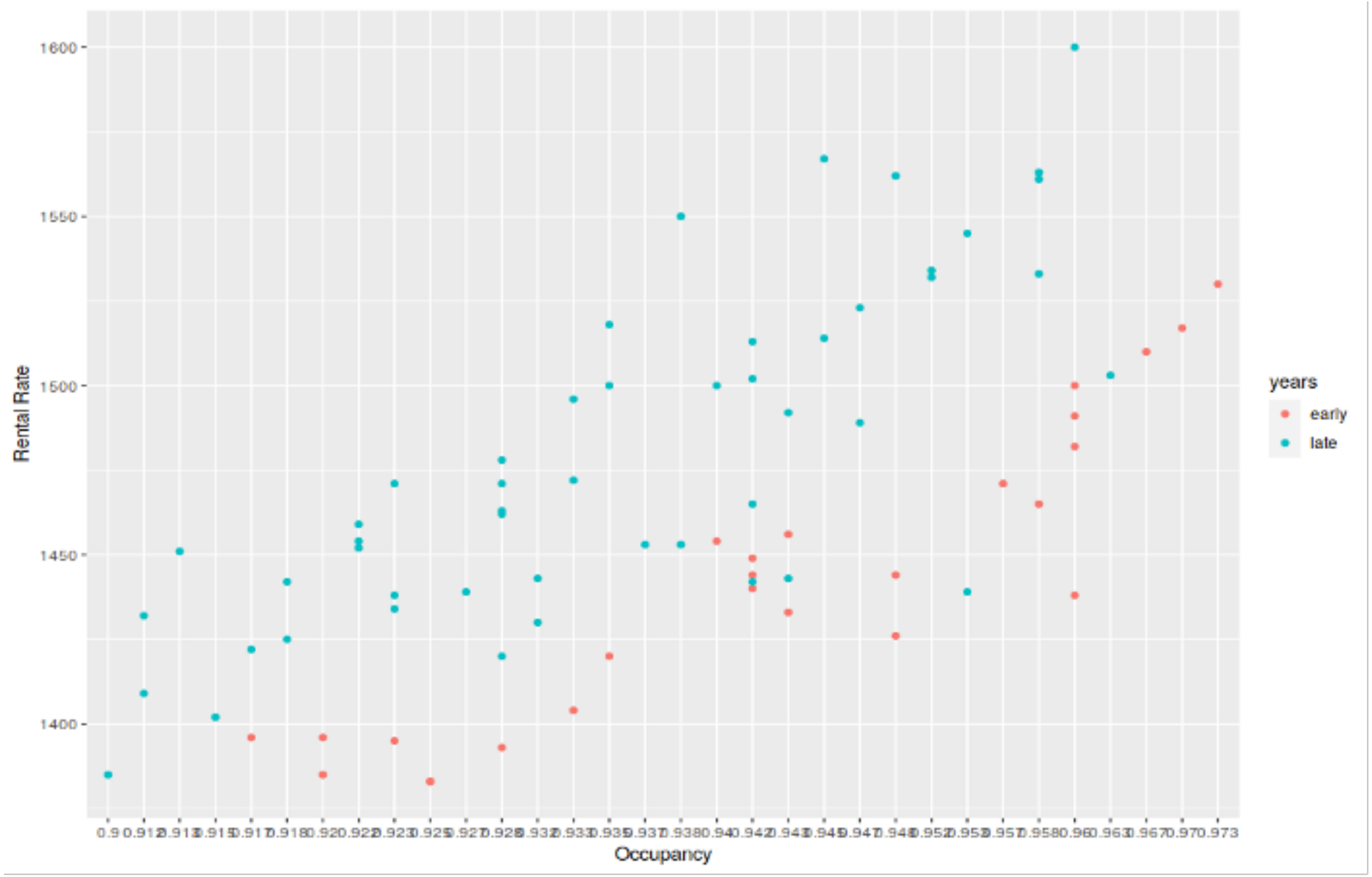
OCCUPANCY FOR STUDIO APARTMENTS

Occupancy

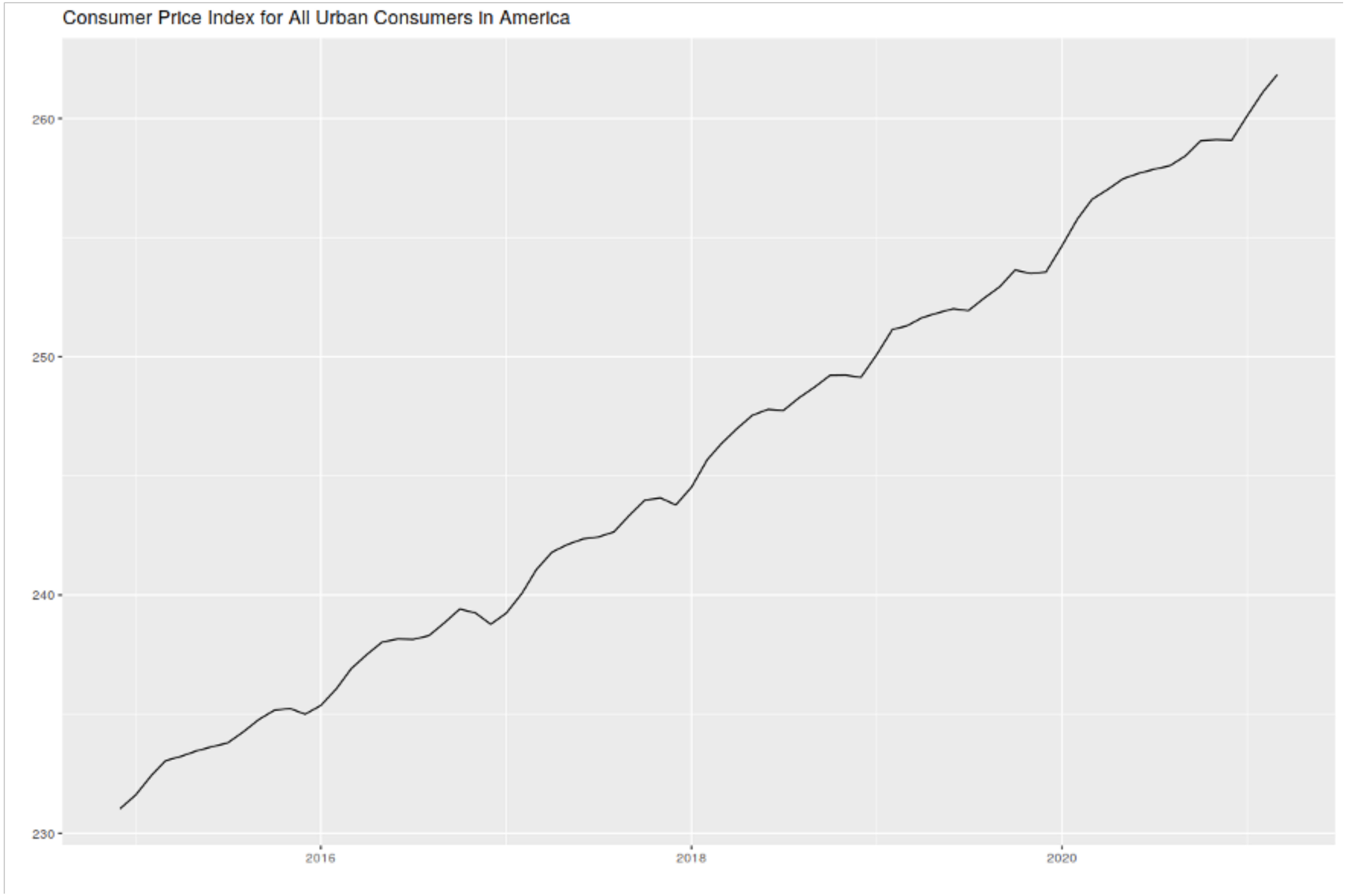
Rental Rate



OCCUPANCY VS RATE



CPI FOR ALL URBAN CONSUMERS IN AMERICA

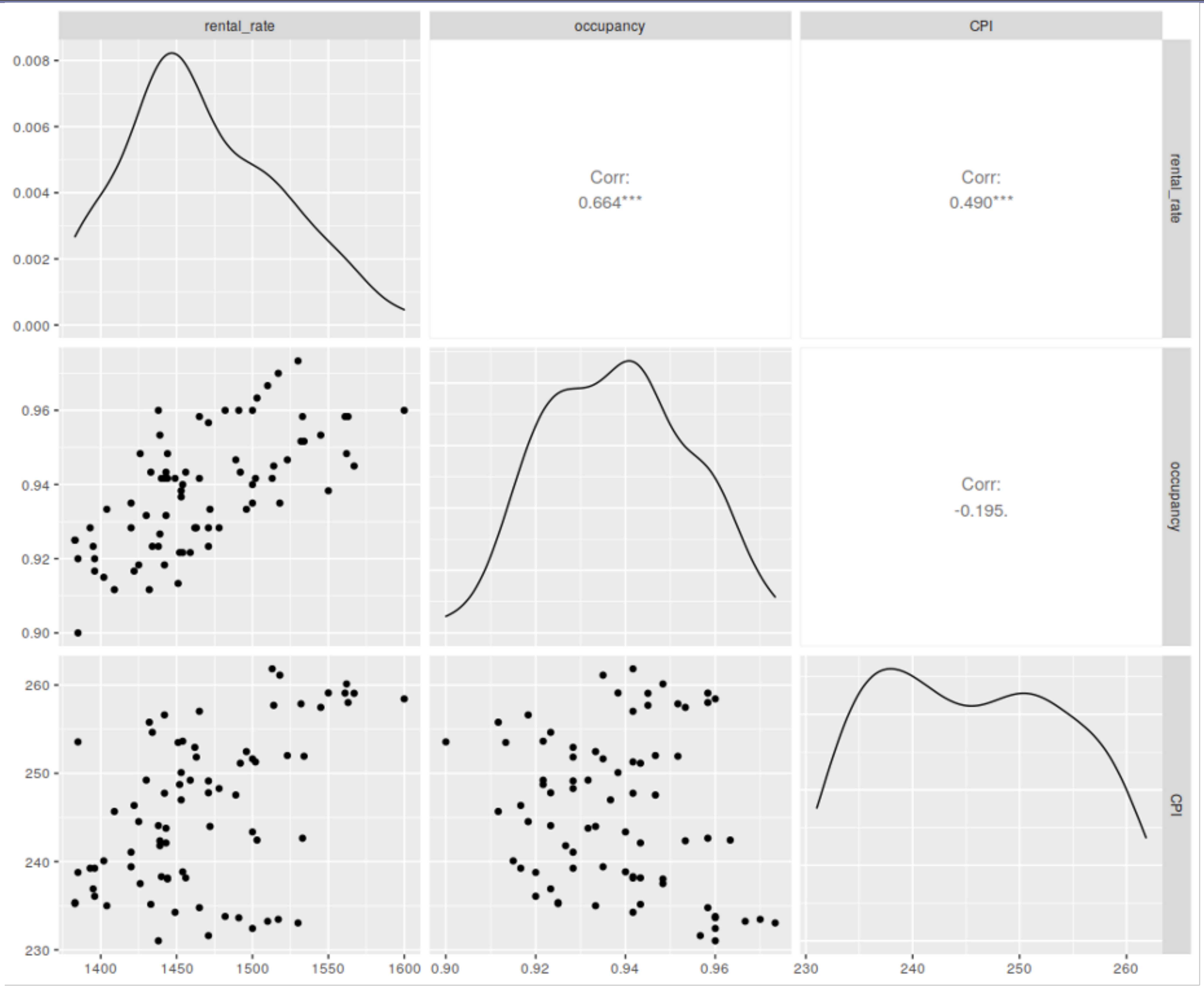


OCCUPANCY, CPI AND RENTAL RATES

Rental Rate

Occupancy

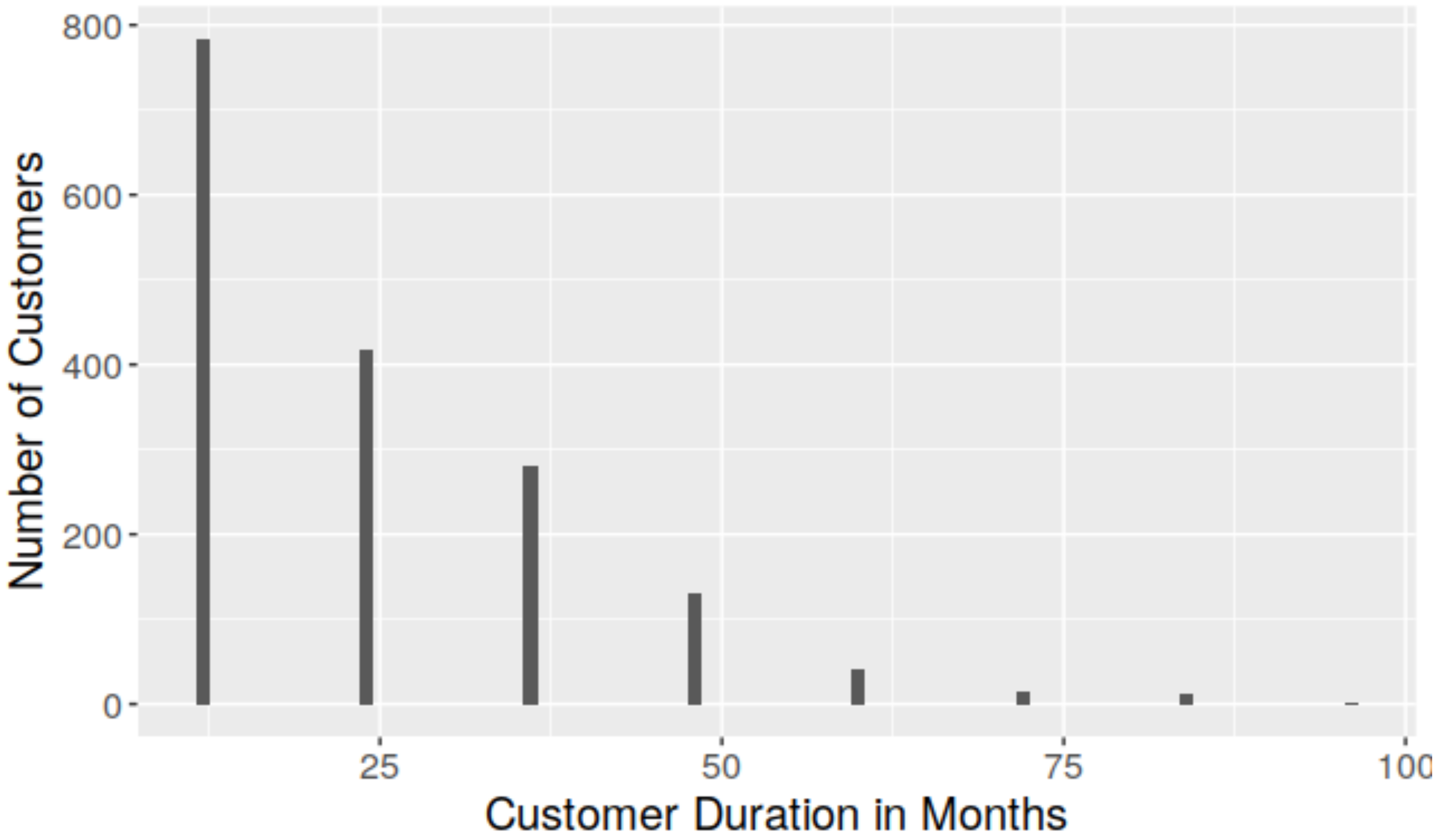
Consumer Price Index



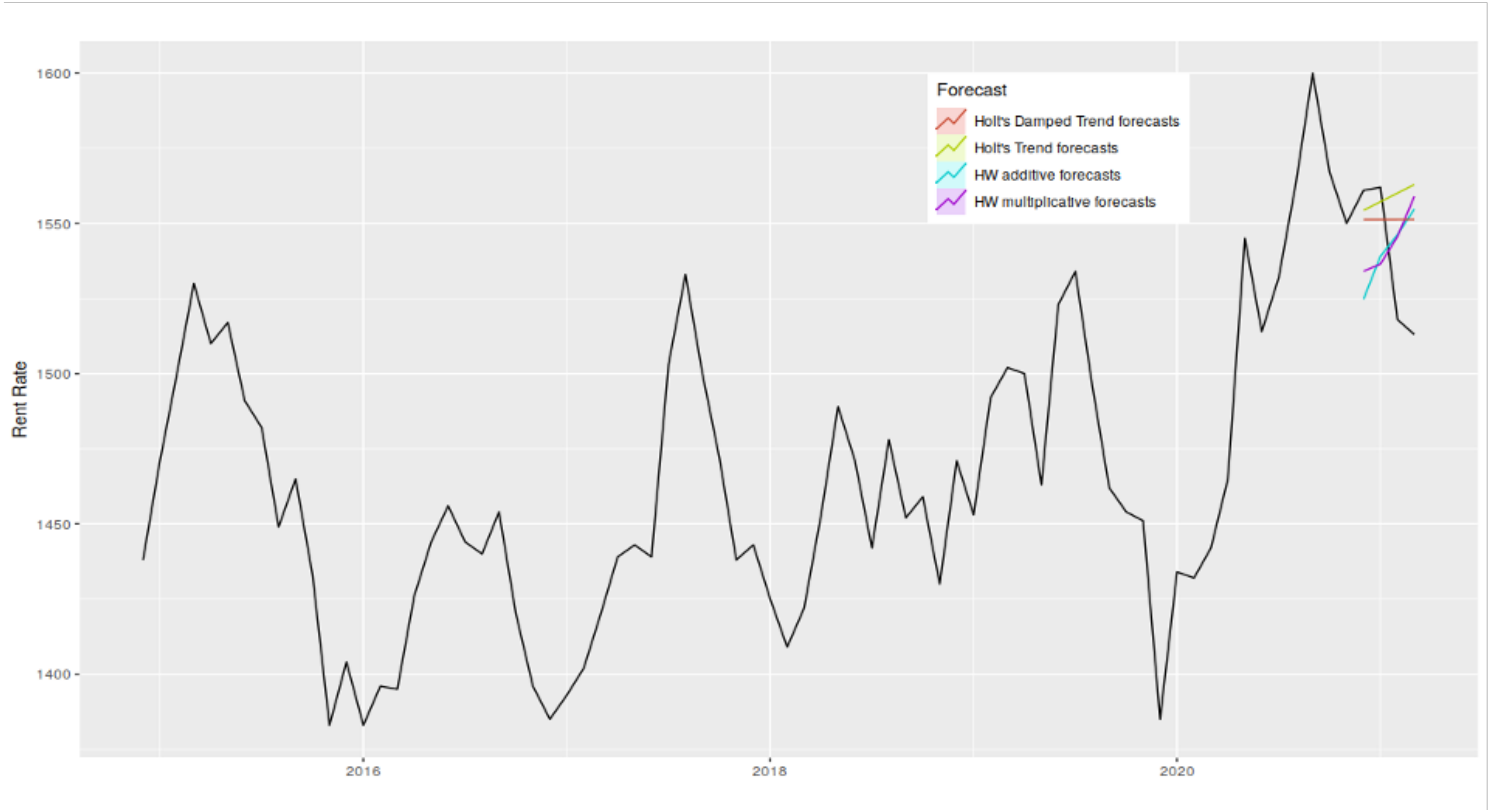
OCCUPANCY BY YEAR

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Occupancy	1022	3507	5795	7155	6972	7015	6964	6975	7058	1746
Rate	102	292	483	596	581	585	580	581	588	582

CUSTOMER DURATION IN MONTHS

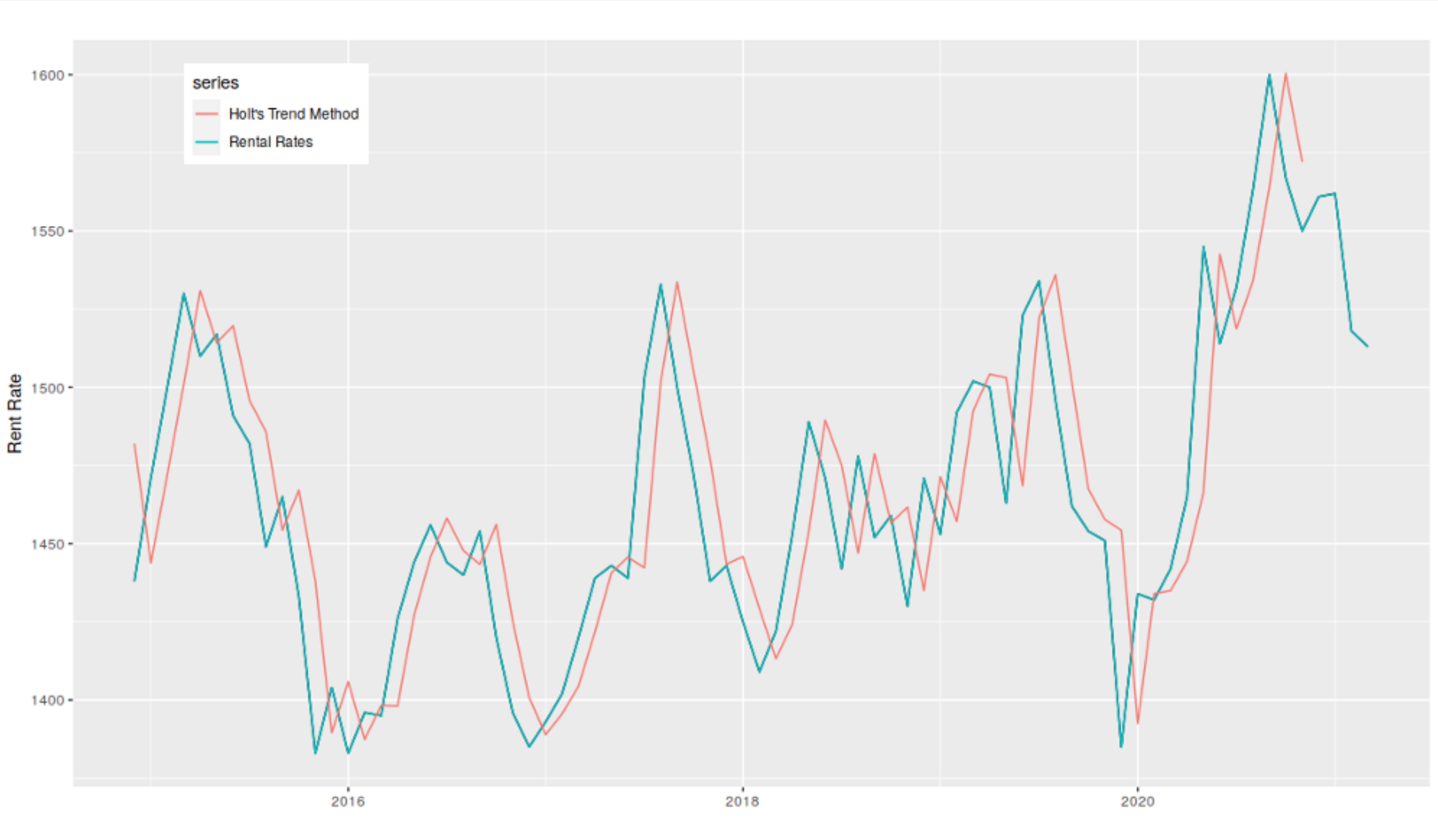


EXPONENTIAL SMOOTHING METHODS

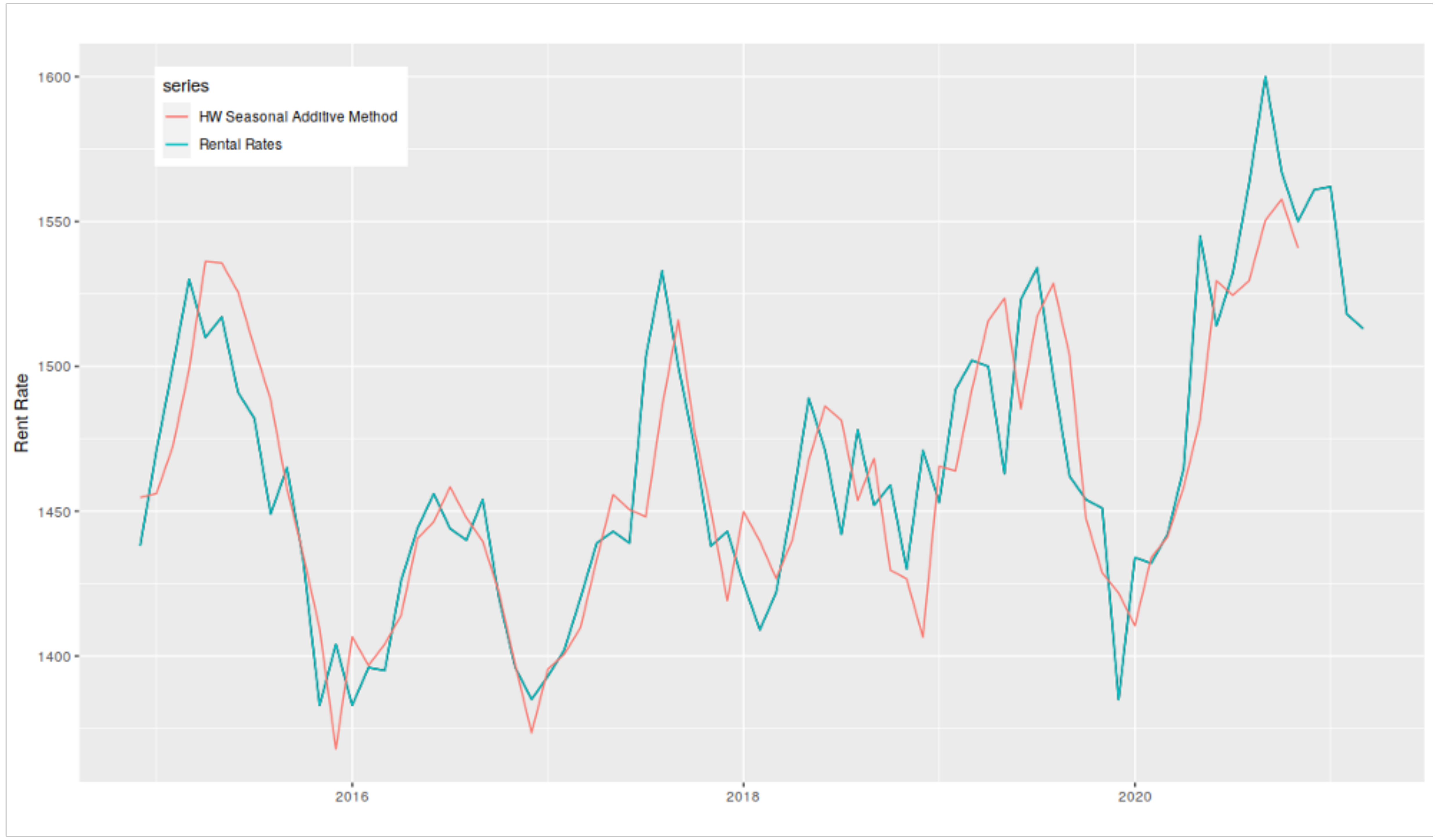


TREND METHOD VS HW SEASONAL ADDITIVE METHOD

Holt's Trend Method

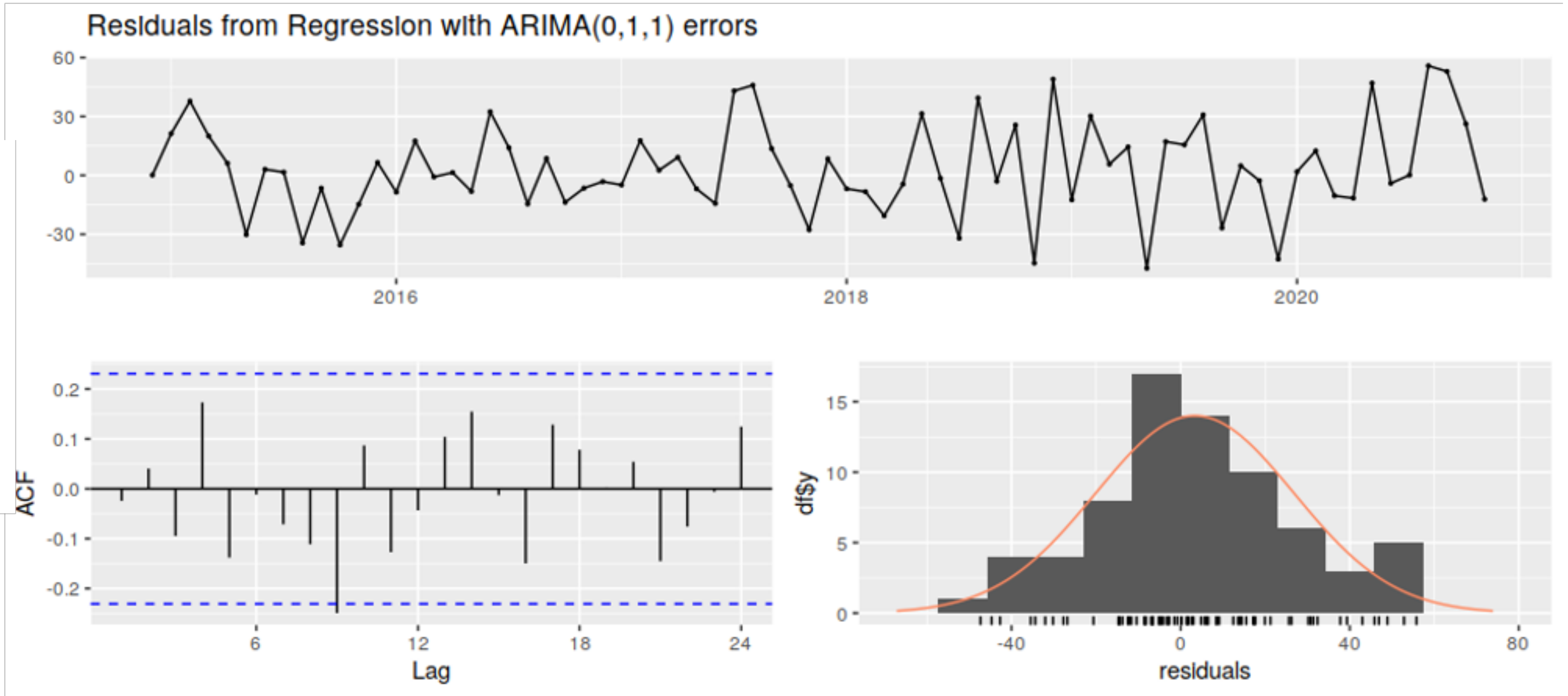


HW Seasonal Additive Method



ARIMA (0, 1, 0) MODEL WITHOUT REGRESSORS

```
ARIMA(0,1,4)                            : 691.709
ARIMA(0,1,4)            with drift      : 695.831
ARIMA(0,1,4)(0,0,1)[12]                 : 693.344
ARIMA(0,1,4)(0,0,1)[12] with drift      : 697.4637
ARIMA(0,1,4)(1,0,0)[12]                 : 692.2382
ARIMA(0,1,4)(1,0,0)[12] with drift      : 696.3634
ARIMA(0,1,5)                            : 693.9006
ARIMA(0,1,5)            with drift      : Inf
ARIMA(1,1,0)                            : 688.071
ARIMA(1,1,0)            with drift      : 692.108
ARIMA(1,1,0)(0,0,1)[12]                 : 690.4376
```



DYNAMIC HARMONIC REGRESSION

In these models, seasonal pattern is modeled using Fourier terms.

These Fourier terms are given to ARIMA models as regressors.

The number of Fourier sin and cos pairs, K (controlling smoothness of the seasonal pattern) is a hyper parameter.

K=1: ARIMA(0,1,1)(1,0,0)[12] gave RMSE = 23.19812 and MAE = 18.29285.

K=2: ARIMA(0,1,1)(0,0,1)[12] gave RMSE = 20.67676 and MAE = 16.45959.

After K=2, models showed lack of fit, therefore, I chose the second model to move forward.

ARIMA MODELS WITH REGRESSORS

ARIMA(0,1,1) with occupancy,

bic = 664.37 , RMSE= 23.58078, MAE = 18.03673

ARIMA(2,0,0) with occupancy and cpi,

bic=677.56, RMSE = 22.33245, MAE = 18.35579

ARIMA(0,1,1) with occupancy and cpi, d is enforced,

bic= 667.25, RMSE= 23.34762, MAE= 18.17315

VECTOR AUTOREGRESSION METHOD

In the earlier models, we assumed that regressors are affecting the series but not vice versa.

However, we have a bidirectional relationship. Occupancy affects the prices and prices affect the occupancy in return.

Vector auto regression models takes this phenomenon into account.

I applied the model with different hyper parameter values, p(number of lags).

The ones which passed the lack of fit tests are the following:

VAR(4) MAE = 40.08055

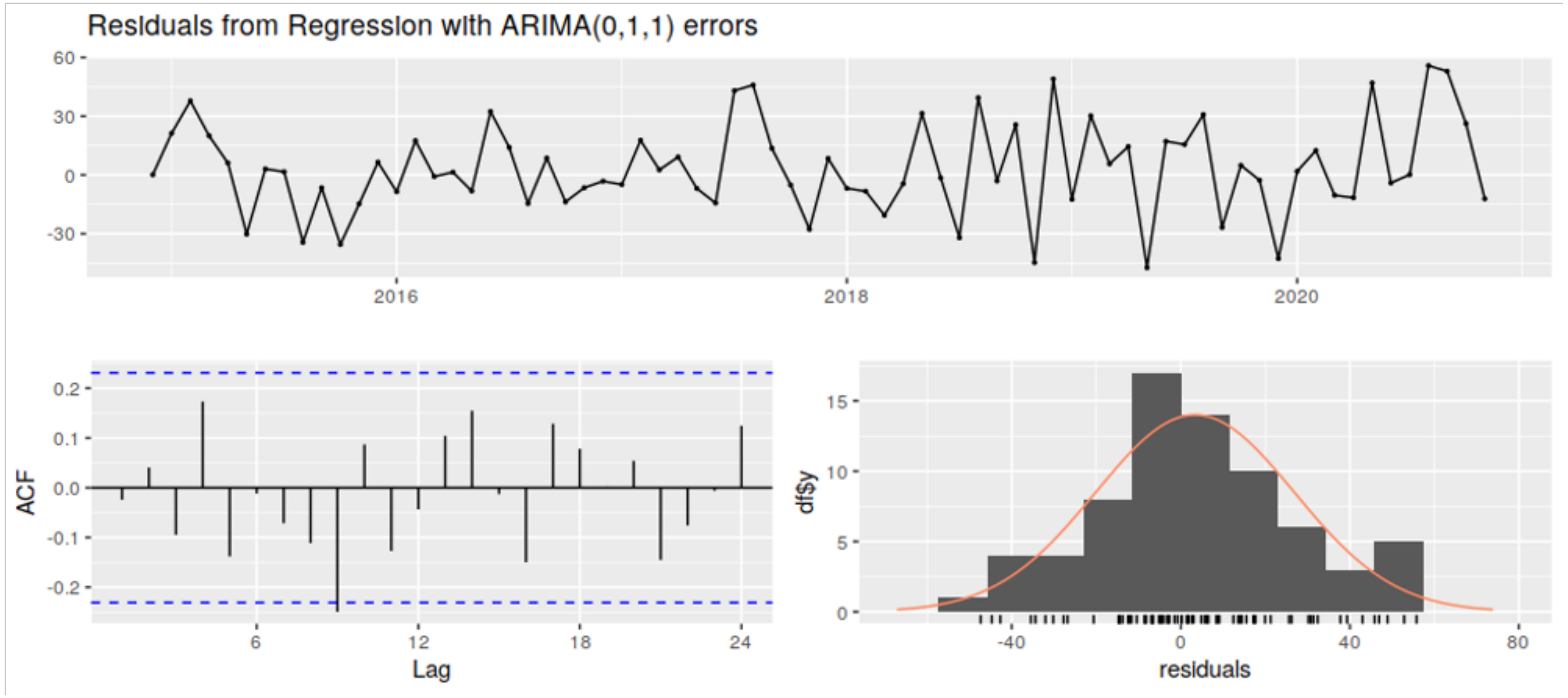
VAR(5) MAE = 40.96711

Since MAE values are significantly higher, I discarded these models.

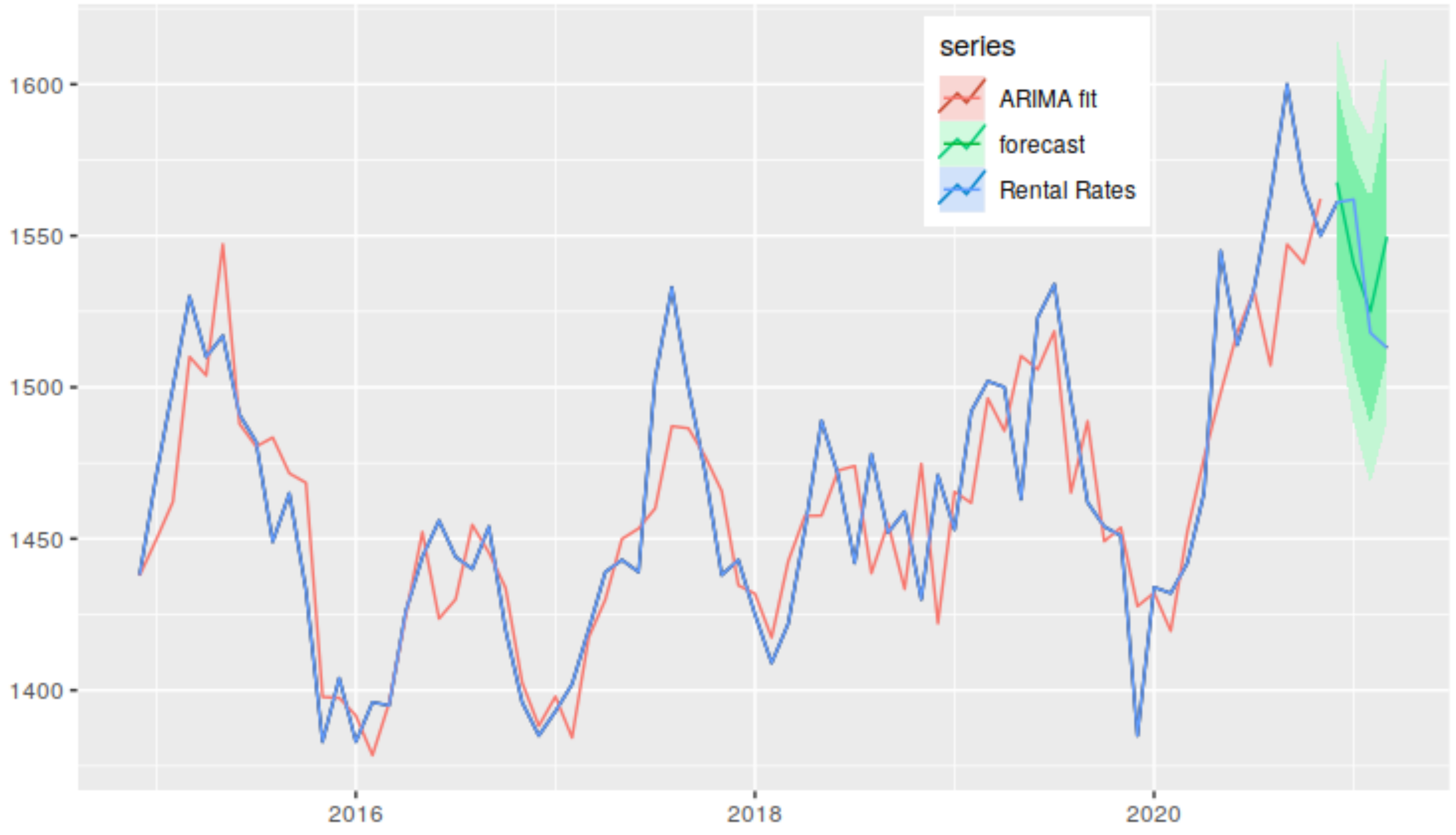
ARIMA(0,1,1) WITH THE REGRESSOR OCCUPANCY

Residuals give a Ljung-Box test p- value of 0.1551 → no lack of fit.

Although dynamic harmonic regression model with K=2 gave smaller accuracy values, we should prefer the model with an external regressor. Let's say we made a bad forecast at some point, external variable coming next month can help us to get back on track. Test MAE value is 17.9.



ARIMA(0,1,1) WITH THE REGRESSOR OCCUPANCY



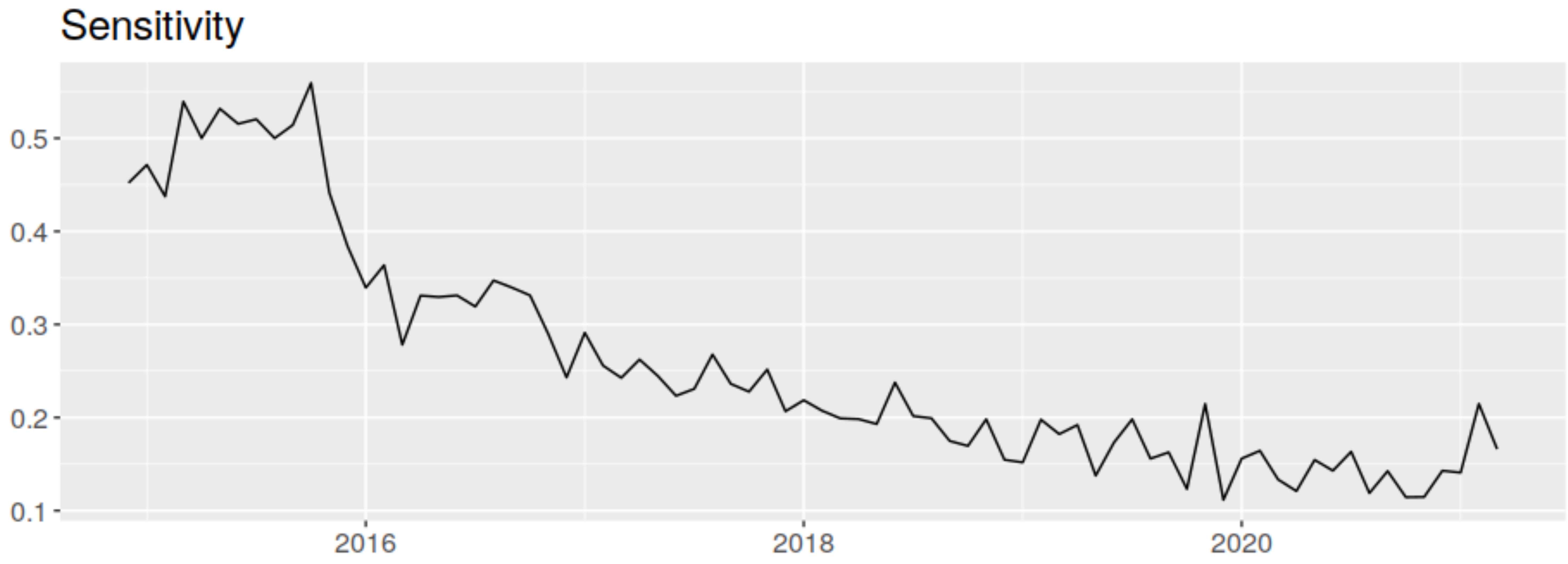
SENSITIVITY

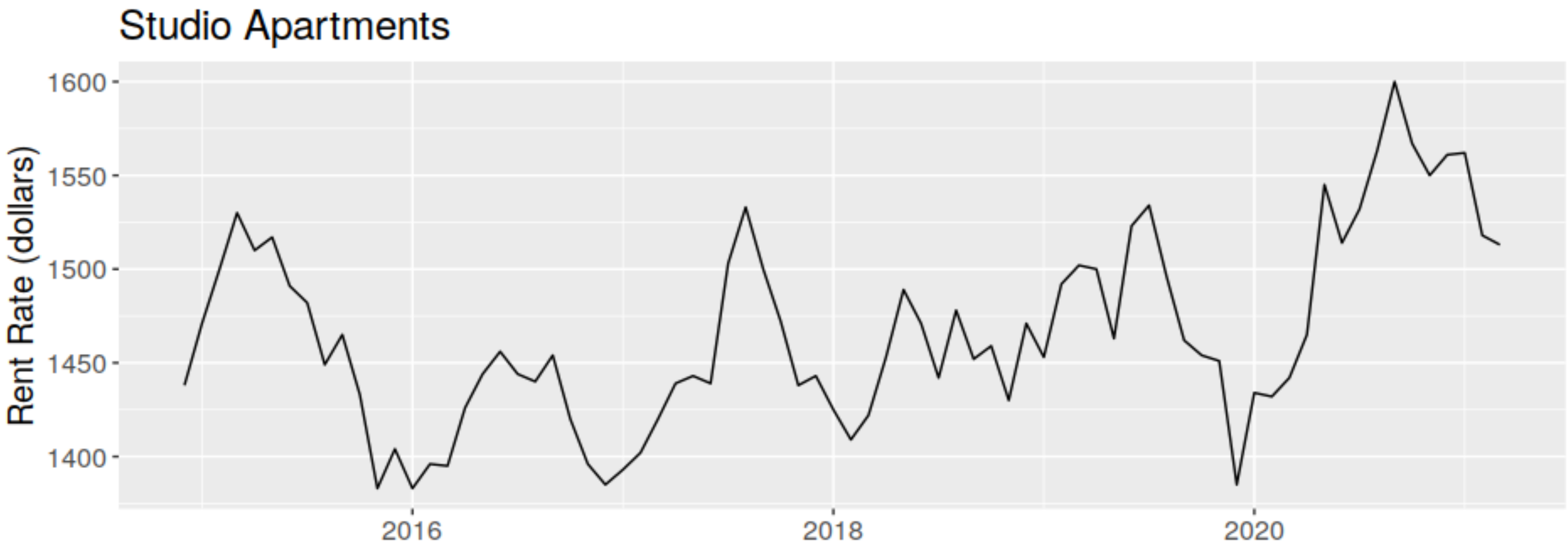
day	month	year	sense	out	possible	gun
2016-11-29	11	2016	0.00000000	0	2	Salı
2016-11-30	11	2016	0.02083333	1	48	Çarşamba
2016-12-01	12	2016	0.62500000	20	32	Perşembe
2016-12-02	12	2016	0.68421053	13	19	Cuma
2016-12-03	12	2016	0.00000000	0	2	Cumartesi
2016-12-04	12	2016	0.00000000	0	2	Pazar
2016-12-05	12	2016	0.00000000	0	1	Pazartesi
2016-12-06	12	2016	0.00000000	0	1	Salı
2016-12-07	12	2016	NaN	0	0	Çarşamba

day	month	year	sense	out	possible	gun
2016-12-29	12	2016	0.00000000	0	1	Perşembe
2016-12-30	12	2016	NaN	0	0	Cuma
2016-12-31	12	2016	0.00000000	0	45	Cumartesi
2017-01-01	1	2017	0.63636364	21	33	Pazar
2017-01-02	1	2017	0.60000000	9	15	Pazartesi
2017-01-03	1	2017	0.00000000	0	4	Salı
2017-01-04	1	2017	0.25000000	1	4	Çarşamba
2017-01-05	1	2017	0.00000000	0	3	Perşembe
2017-01-06	1	2017	0.50000000	2	4	Cuma

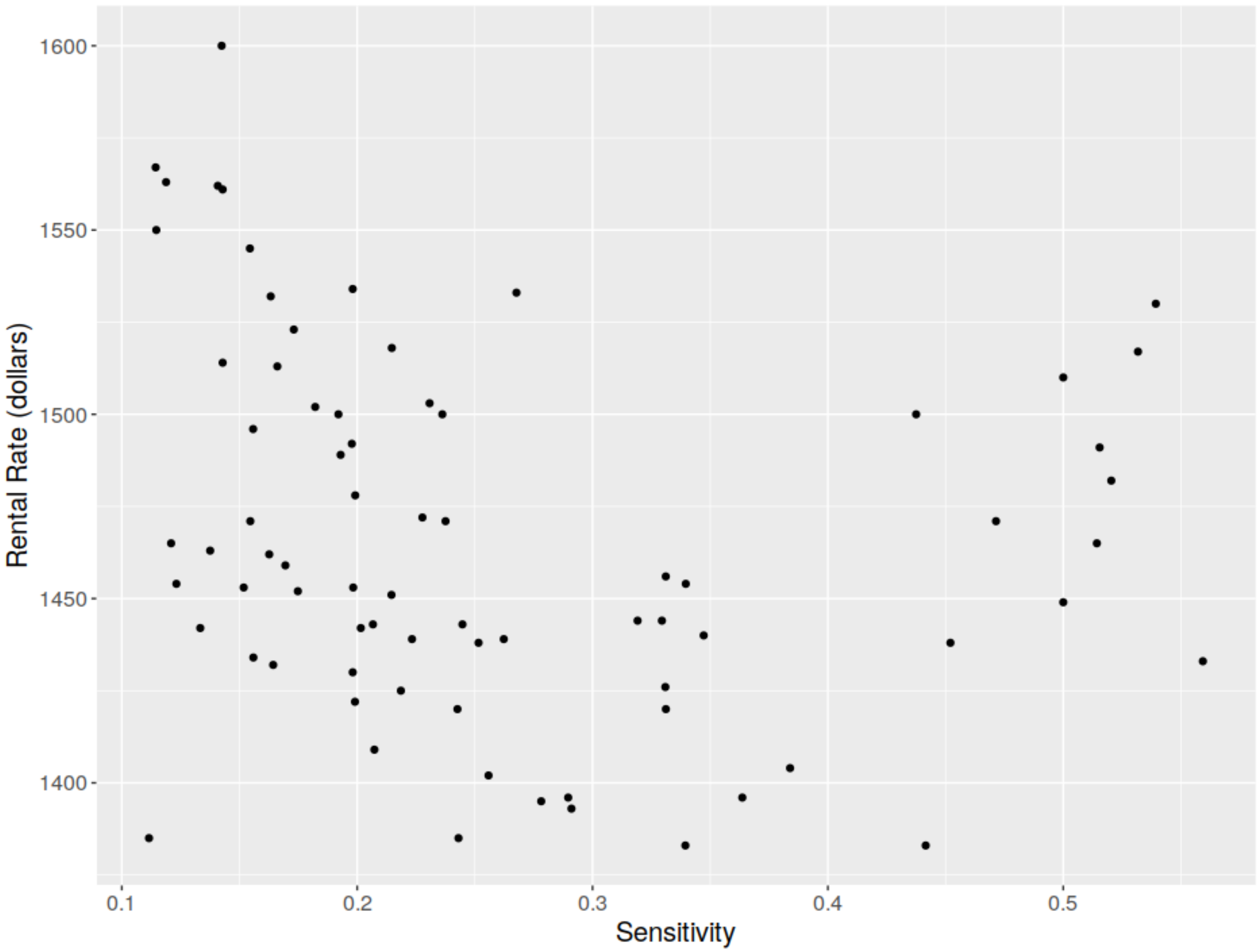
day	month	year	sense	out	possible	gun
2016-09-27	9	2016	0.40000000	2	5	Salı
2016-09-28	9	2016	0.00000000	0	4	Çarşamba
2016-09-29	9	2016	0.16666667	1	6	Perşembe
2016-09-30	9	2016	0.10714286	3	28	Cuma
2016-10-01	10	2016	0.55000000	22	40	Cumartesi
2016-10-02	10	2016	0.00000000	0	2	Pazar
2016-10-03	10	2016	0.66666667	4	6	Pazartesi
2016-10-04	10	2016	NaN	0	0	Salı
2016-10-05	10	2016	1.00000000	1	1	Çarşamba
2016-10-06	10	2016	0.00000000	0	1	Perşembe

SENSITIVITY

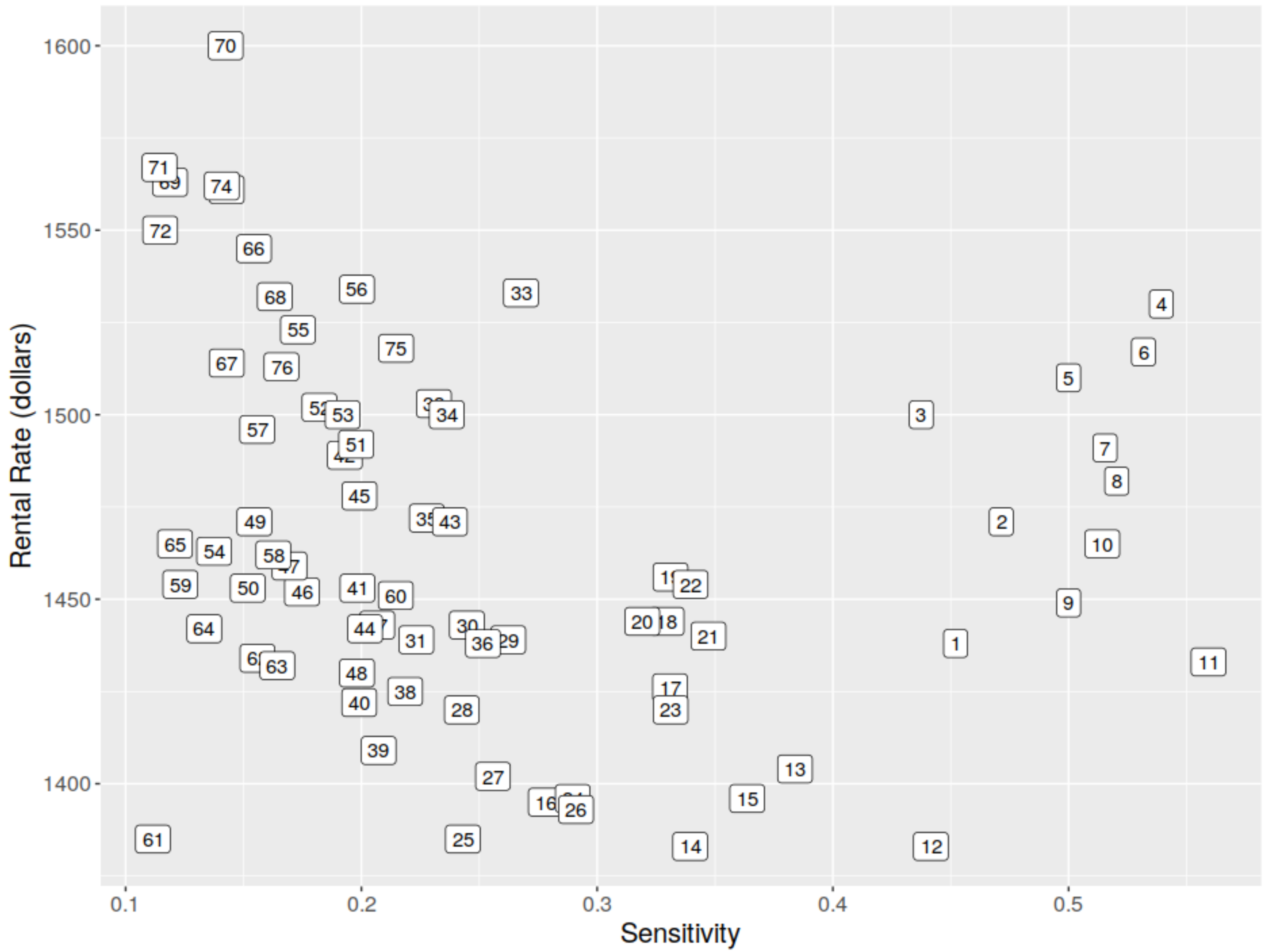




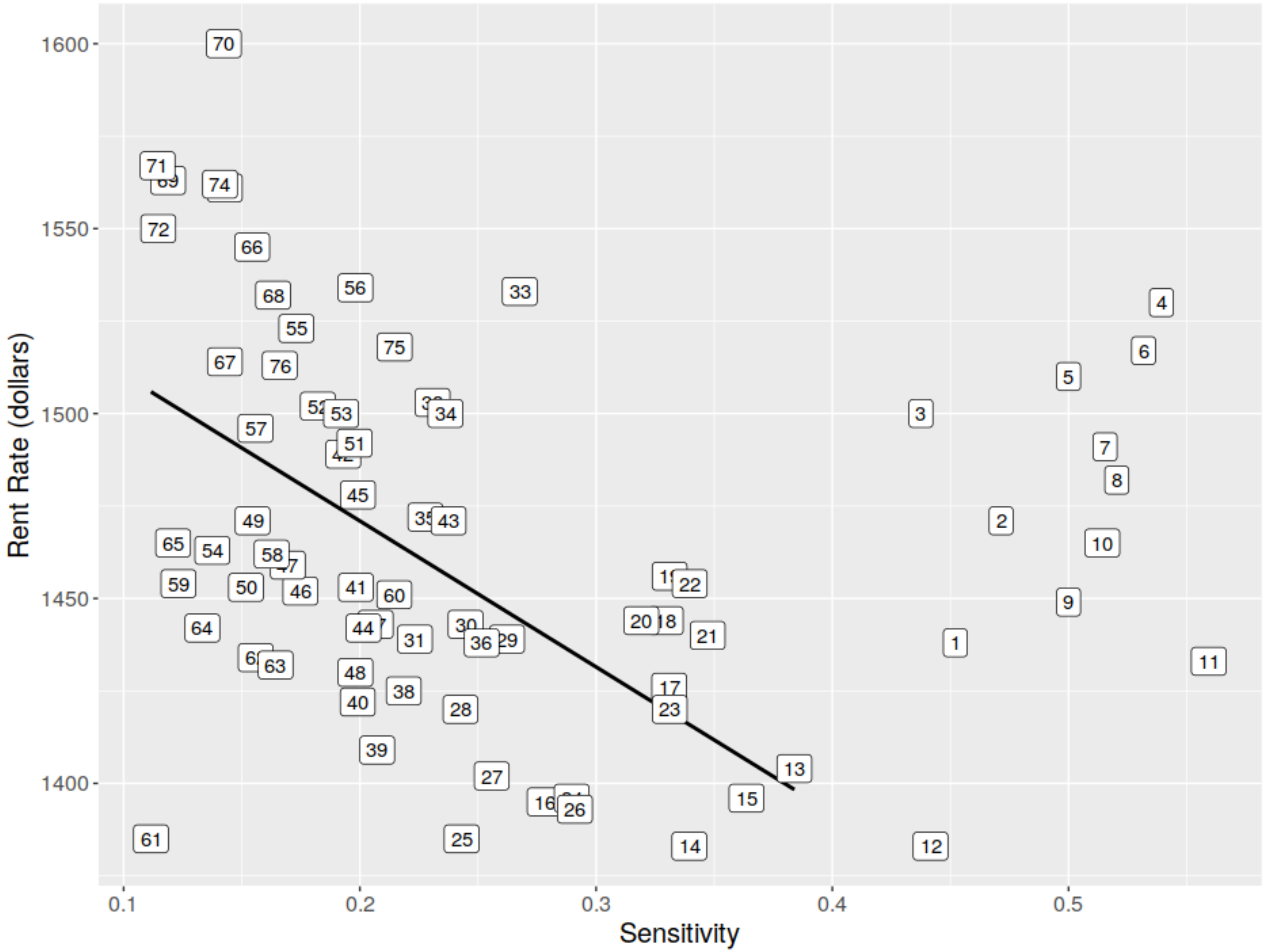
SENSITIVITY



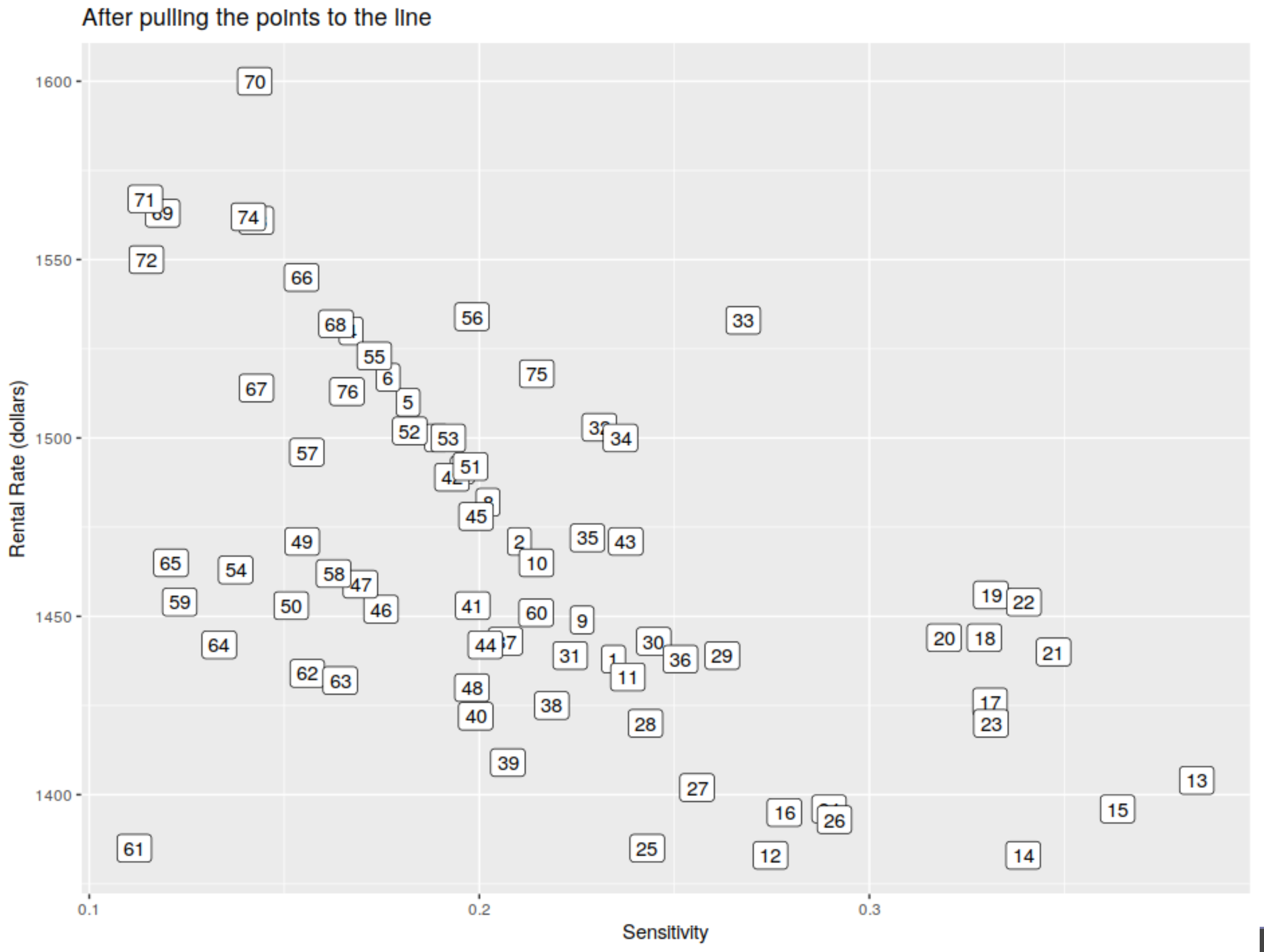
SENSITIVITY



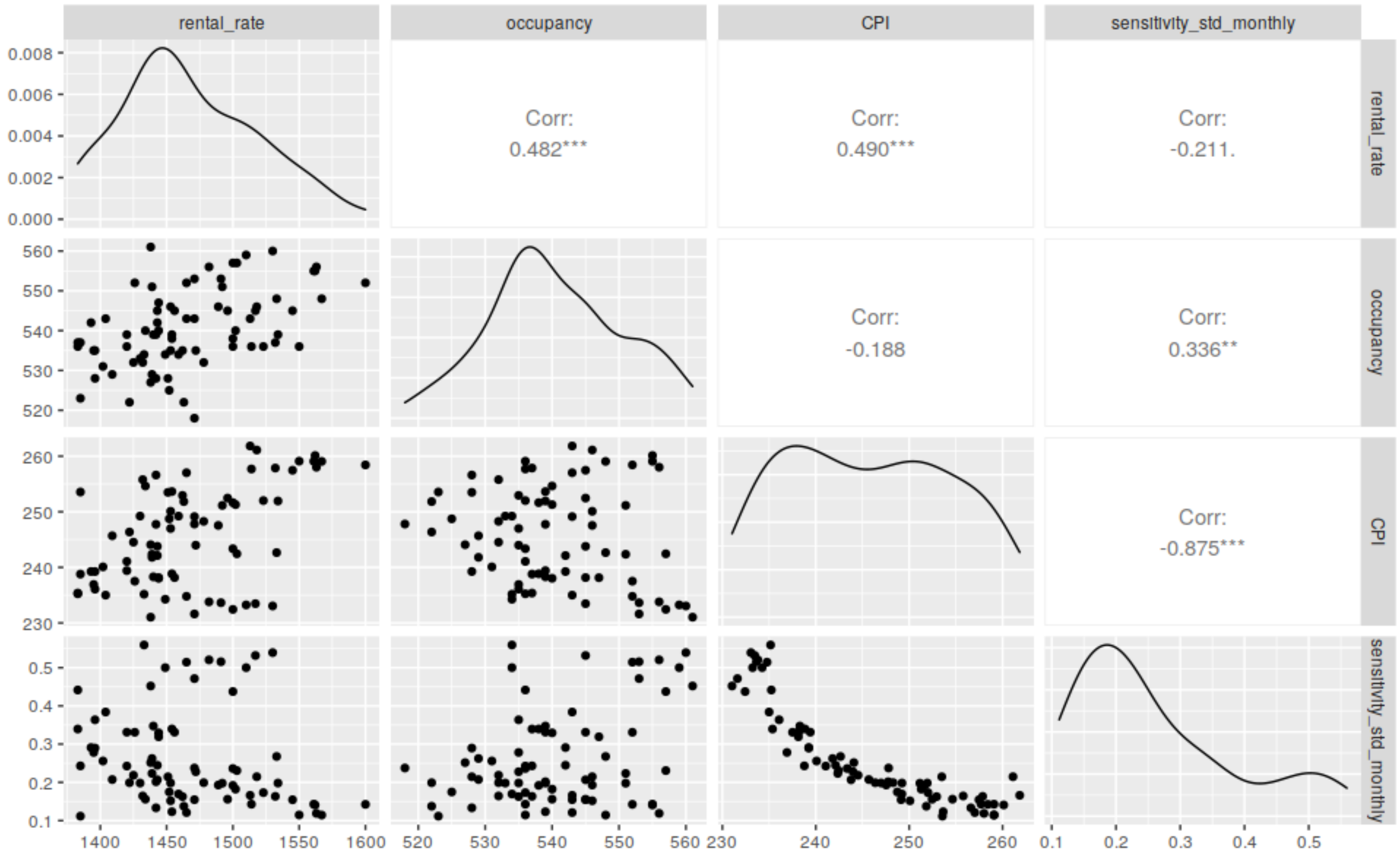
SENSITIVITY



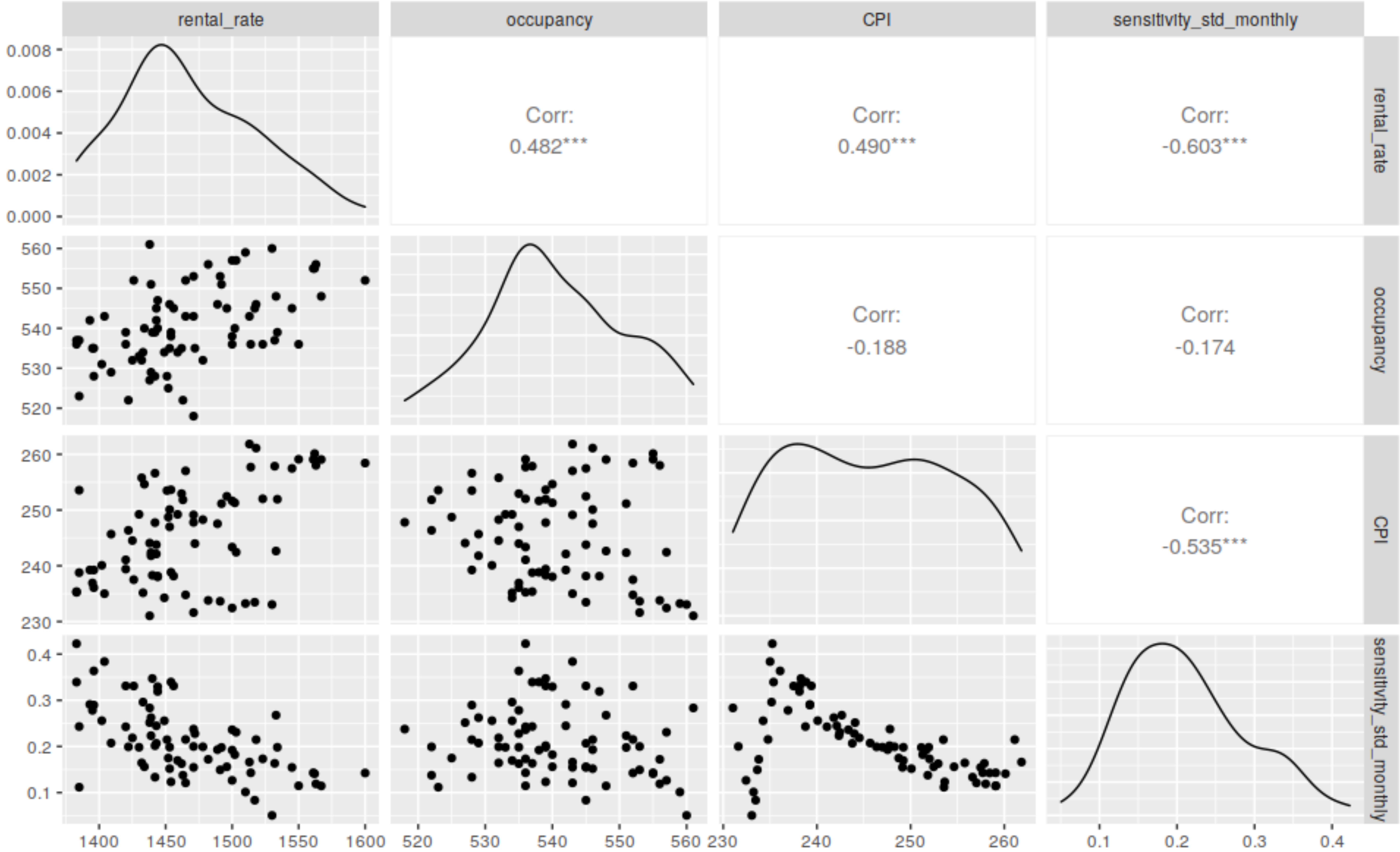
SENSITIVITY



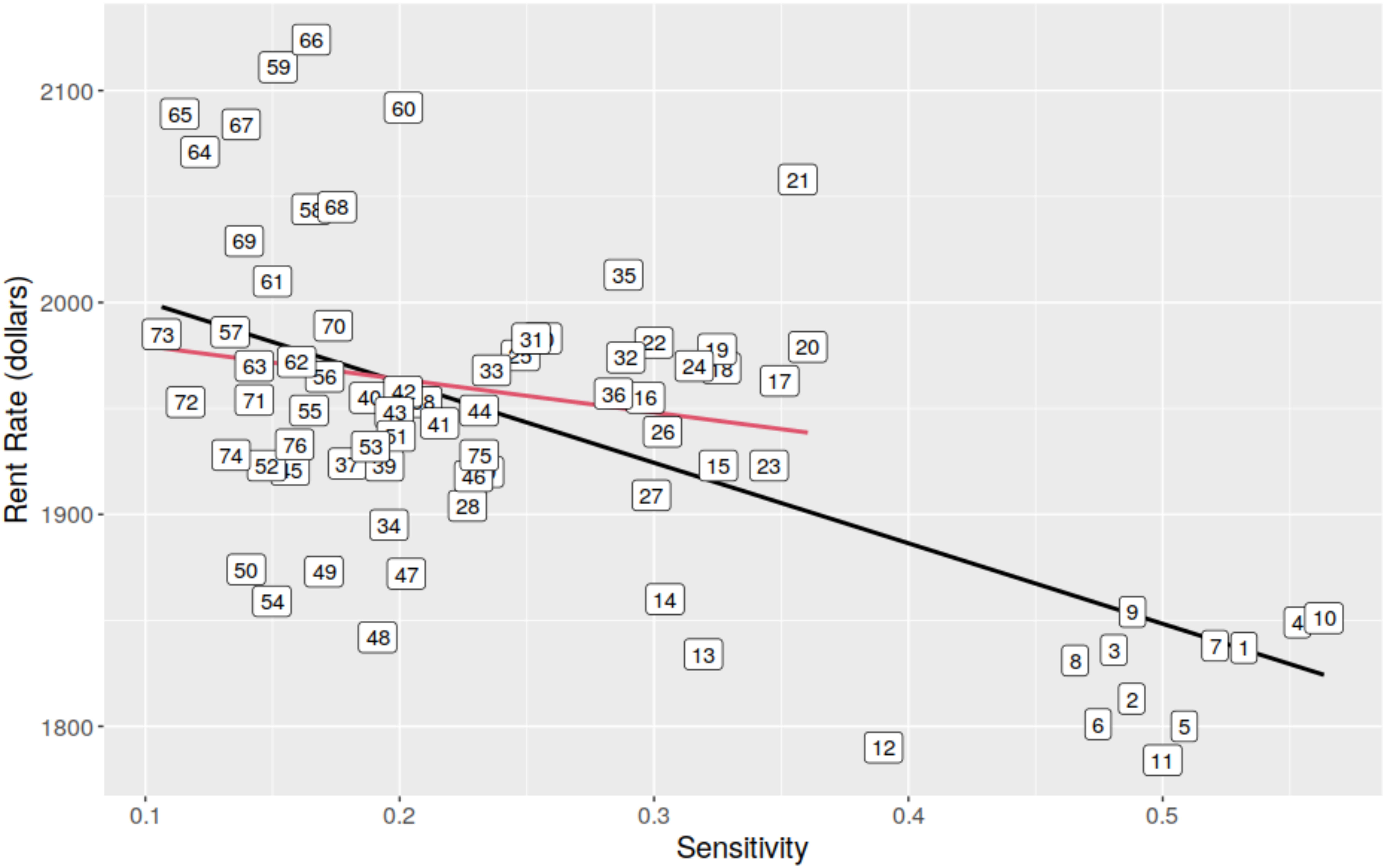
SENSITIVITY



SENSITIVITY



SENSITIVITY
ONE BEDROOM APARTMENTS



MODELS FOR ONE BEDROOM APARTMENTS

MODEL	REGRESSORS	LACK OF FIT	MAE
ARIMA (0,1,0)	ALL REGRESSSORS	✓	31.58
ARIMA (0,1,0) (0,0,1) 12	OCCUPANCY, CPI	✓	30.19
ARIMA (0,1,0)	OCCUPANCY, SENSITIVITY	✓	31.71
ARIMA (0,1,0)	CPI, SENSITIVITY	✓	31.59
ARIMA (0,1,0) (0,0,1) 12	OCCUPANCY	✓	30.73
ARIMA (0,1,0)	SENSITIVITY	✓	31.76
DYNAMIC HARMONIC REGRESSION	K=1, 2, 3, 4, 5, 6	LACK OF FIT	-
VAR(1), VAR(2), VAR(3)	ALL REGRESSORS	NOT STABLE	-
VAR(1)	OCCUPANCY, SENSITIVITY	✓	26.92
VAR(2)	OCCUPANCY, SENSITIVITY	✓	26.13
VAR(3)	OCCUPANCY, SENSITIVITY	✓	25.28
VAR(1)	OCCUPANCY	✓	29.15

FUTURE IMPROVEMENTS

- To avoid overfitting we can apply cross validation on the training set. But when we have a short time series, it might make us lose important information and make the forecasts meaningless.
- Instead of using only one model, we can use a combination of them. Utilizing the wisdom of the crowd might help us with both overfitting and biasedness.
- We can apply non-linear time series models such as STAR, ESTAR or LSTAR.
- When we don’t have enough data, we can use bootstrapping and bagging to get better results.
- We can compare the models according to how they behave with different horizon lengths. We might choose according to our needs.

FUTURE IMPROVEMENTS

- We can add competitor rates, demographics, economic indicators, sales channels, promotions, length of the stay, and various other variables into the models as regressors.
- Instead of segregating the data at only UnitType level, we can segregate according to the sensitivity of the customers. This might help us to not lose sensitive customers, and gain more profit from the customers with low sensitivity. But we need more data for that.
- Outlier detection is important. Model should be able to detect sudden changes so that we can gather more information. Is a park or metro station recently added to the neighborhood? Or there was a sudden drop in the occupancy because of a natural disaster?

FUTURE IMPROVEMENTS

- We can detect structural breaks. We can utilize this information by creating spline like models instead of getting rid of the earlier chunk and losing valuable information.
- We can use lagged regressors instead of only the regressors themselves since the effect can be lagged. For example, rent rate of today might be affected by occupancy of yesterday instead of today.
- External variables are important, when we have to decide between the models with and without external variables where accuracy values are similar, we might choose the one with the external variables. Let’s say we made a bad forecast at some point, external variable coming next month can help us to get back on track.

THANK YOU FOR YOUR ATTENTION

ANY QUESTIONS?