STATISTICAL PROJECT

Arianna Morandi & Beatrice Spagnolo & Tomas Guarini

Overview

1. Obtaining data
2. Clean and filter data
3. Explore data
4. Model data
5. Interpreting data

Flights data

Case study

Analysis of delays of flights arriving in NY

Datasets:

- US_flights_2023
- weather_meteo_by_airport

Clean and filter data

Steps

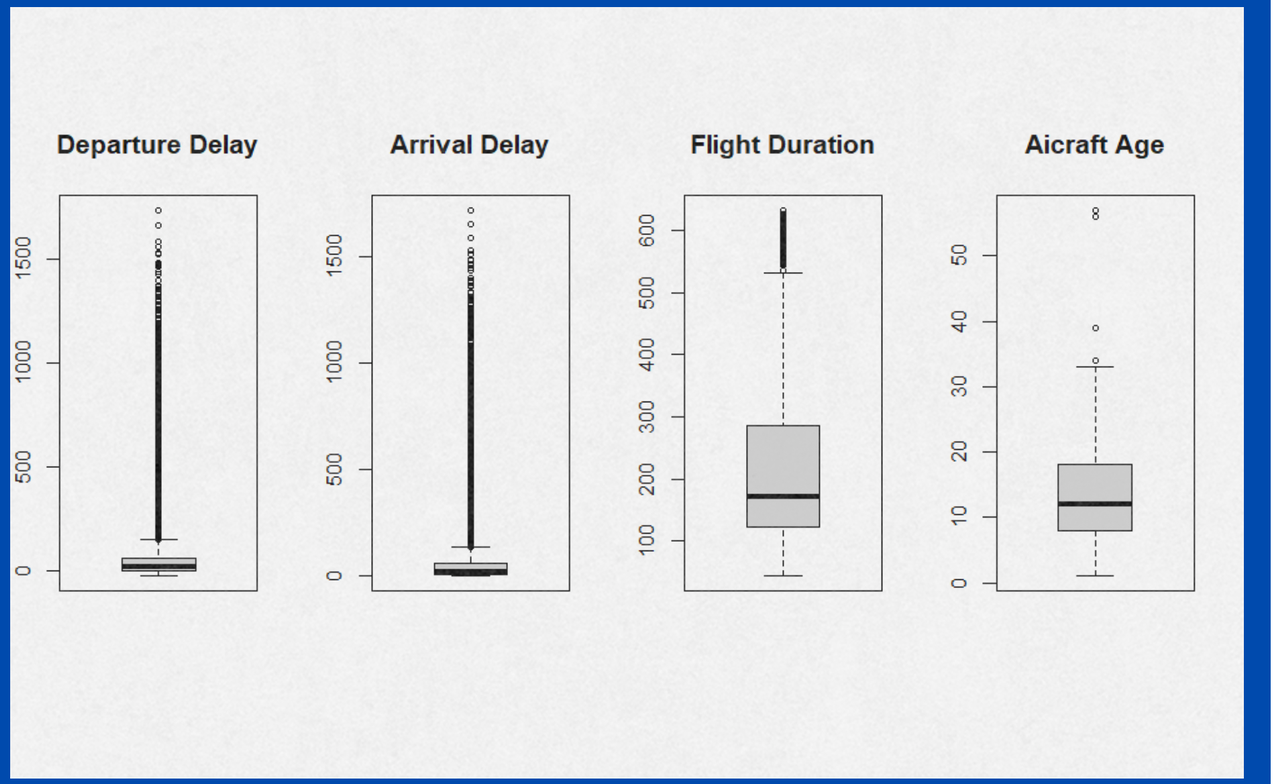
- selecting flights that arrived in the JFK airport of NY
- merge of flights and weather datasets
- omit na values
- delete non useful variables
- select only delayed flights
- create a categorical variable for the moments of the day

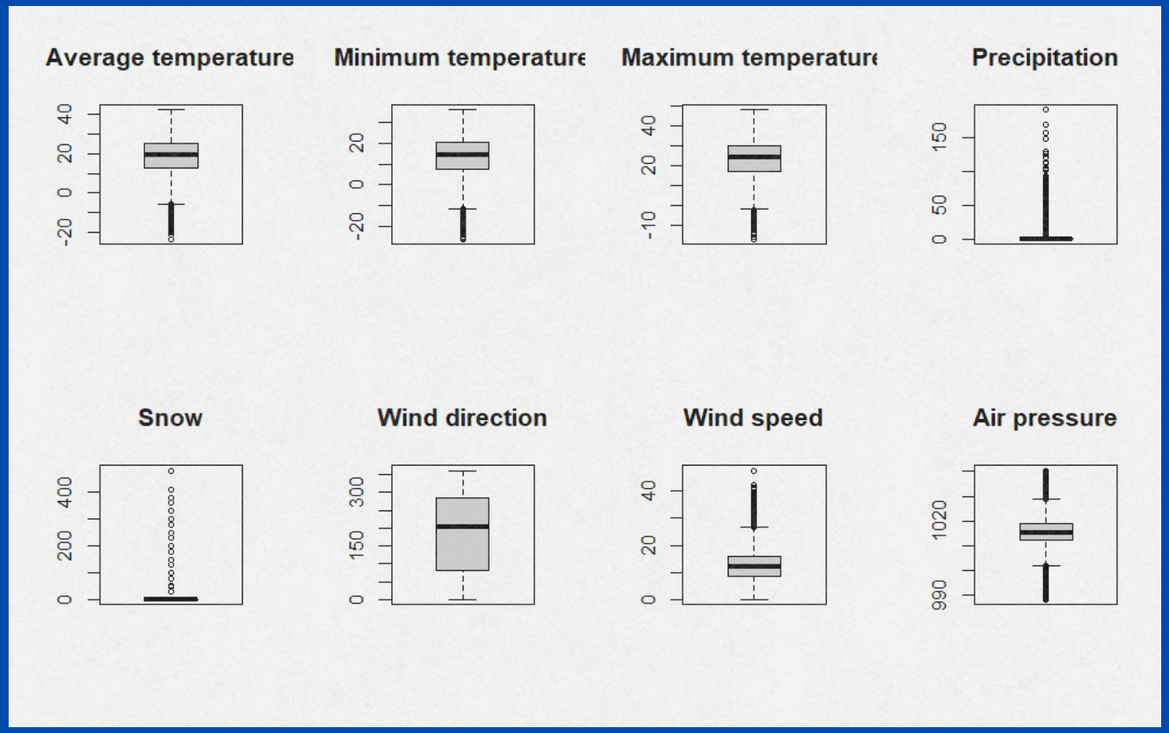
Variables of interest

Arr_Delay, Dep_Delay, Flight_Duration, Aicraft_age, tavg, tmin, tmax, prcp, snow, wdir, wspd

Explore data

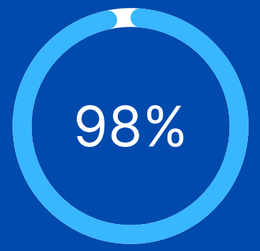
- boxplots of all flights and weather variables after outlier exclusion





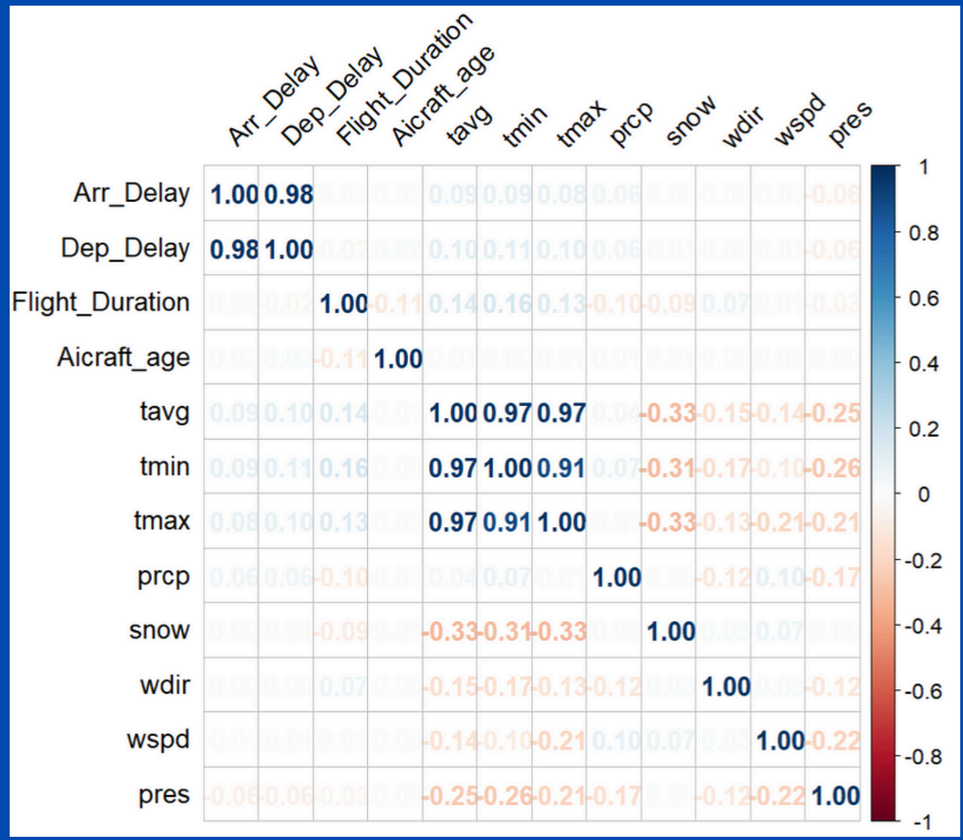
Explore data

- covariance and correlation matrices



Dep_Delay

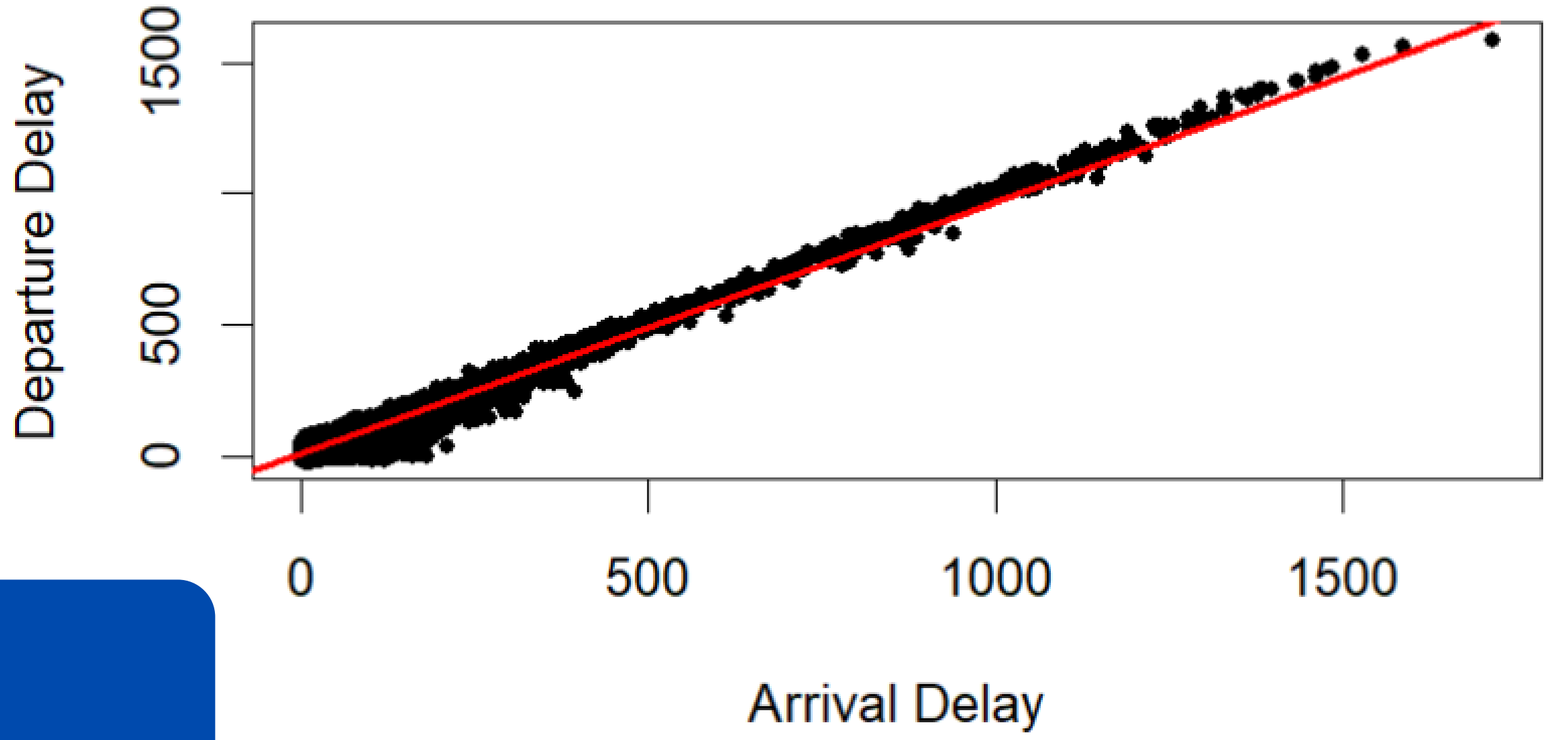
"Dep_Delay" is the most correlated variable, while all the others are close to zero.



Arr_Delay	Dep_Delay	Flight_Duration	Aicraft_age
10988.279530	11076.286379	116.883910	13.257958
tavg	tmin	tmax	prcp
84.116435	89.186819	82.101217	58.659624
snow	wdir	wspd	pres
-3.410022	19.125641	-4.745323	-38.732823

Model data

REGRESSION



Train and test split

Train from October to December
dataset$FlightDate < "2023-09-30"

Simple linear regression

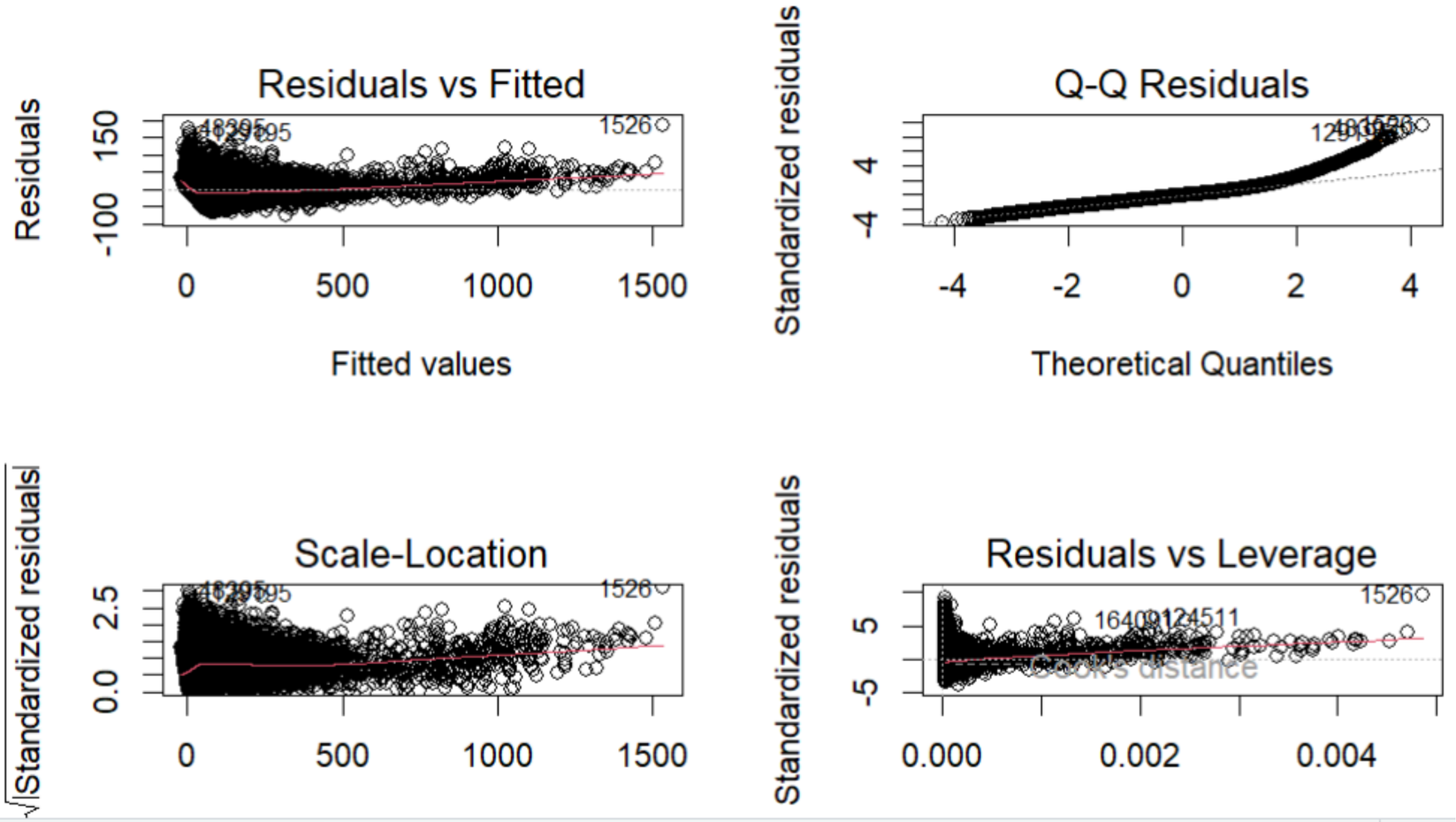
- **Model**
- **Predictions**
- Plot
- Outlier
- Non-linear transformations

Arr_Delay ~ Dep_Delay

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.7919020  0.1072867   63.31   <2e-16 ***
Dep_Delay   0.9613715  0.0008784 1094.50   <2e-16 ***

Residual standard error: 19.37
on 39920 degrees of freedom

Multiple R-squared:  0.9678
Adjusted R-squared:  0.9677

                          fit      lwr      upr
Prediction interval: 1 16.42575 16.22244 16.62906
                          fit       lwr      upr
Confidence interval: 1 16.42575 -21.48971 54.34121
```

Simple linear regression

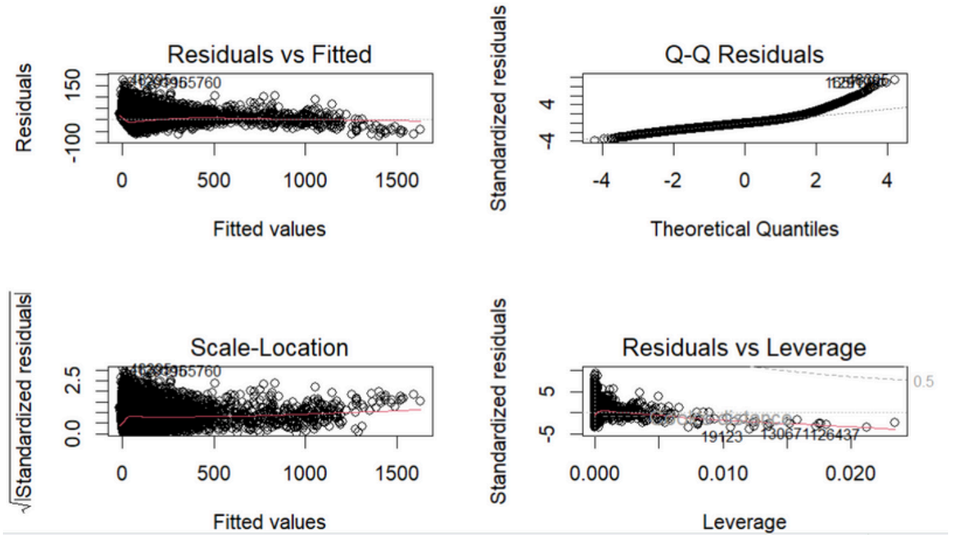
- Model
- Predictions
- **Plot**
- **Outlier**
- Non-linear transformations



Outlier: 1526 R-squared: 0.9676

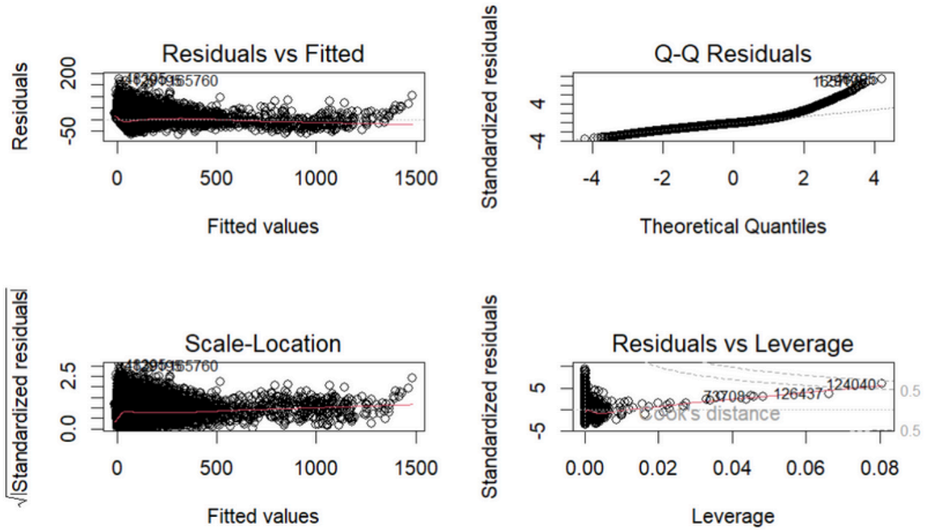
Non-linear transformation

Polynomial 2nd degree



R-squared: 0.9694

Polynomial 3rd degree



R-squared 0.9702

log(Arr_Delay) ~ Dep_Delay

R-squared 0.40324

Multiple linear regression

- Backward stepwise procedure
- Aic and Bic
- Final model
- Plots
- Collinearity

Arr_Delay ~ Dep_Delay + Flight_Duration + Aicraft_age + tavg + tmin + tmax + prcp + snow + wdir + wspd + pres + DepTime_categoric

Coefficients initial model

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.947e+01	1.906e+01	2.071	0.03833	*
Dep_Delay	9.635e-01	8.723e-04	1104.533	< 2e-16	***
Flight_Duration	3.602e-02	1.044e-03	34.499	< 2e-16	***
Aicraft_age	-4.430e-02	1.347e-02	-3.288	0.00101	**
tavg	2.680e-01	1.022e-01	2.622	0.00873	**
tmin	-3.496e-01	5.682e-02	-6.153	7.70e-10	***
tmax	-1.139e-01	5.657e-02	-2.014	0.04400	*
prcp	5.250e-02	1.011e-02	5.190	2.11e-07	***
snow	2.577e-02	3.962e-03	6.503	7.95e-11	***
wdir	9.295e-05	9.309e-04	0.100	0.92047	
wspd	-1.954e-02	1.819e-02	-1.074	0.28263	
pres	-3.596e-02	1.861e-02	-1.932	0.05332	.
DepTime_categoric2	8.878e-01	2.308e-01	3.847	0.00012	***
DepTime_categoric3	-2.385e+00	2.526e-01	-9.439	< 2e-16	***
DepTime_categoric4	1.059e+00	6.713e-01	1.578	0.11468	

Adjusted R-squared: 0.969

Backward stepwise procedure #1

excluding: wind direction, wind speed, air pressure, maximum temperature

```
mod4: R2 0.9689894 AIC 348126.1 BIC 348229.3
```

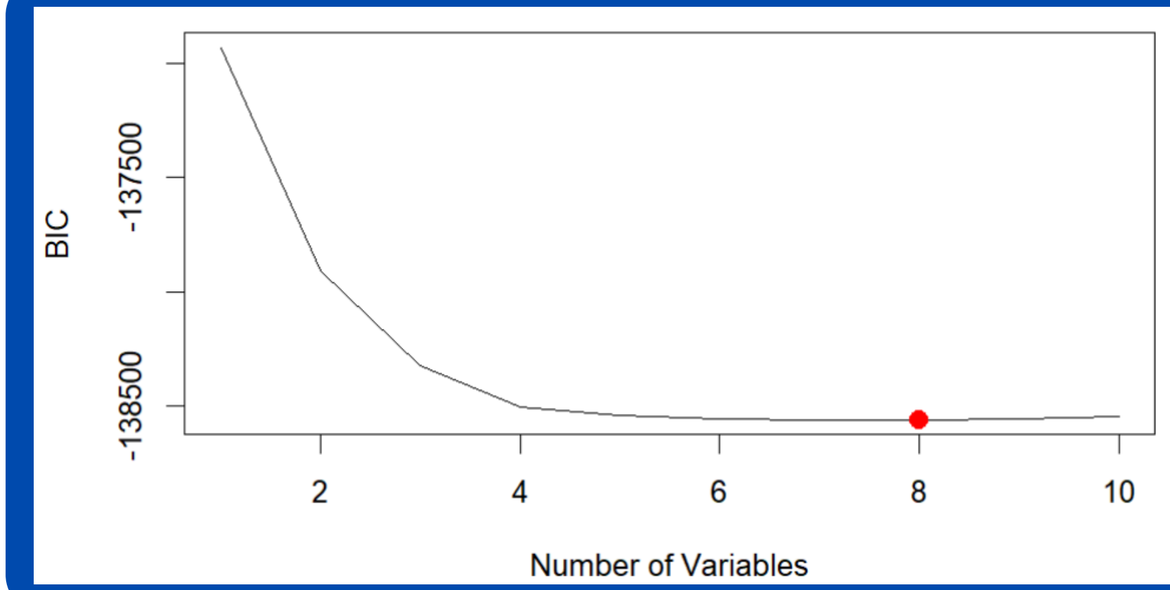
excluding: average temperature

```
mod5: R2 0.968986 AIC 348128.5 BIC 348223
```

excluding: aircraft age

```
mod6: R2 0.9689808 AIC 348135.2 BIC 348229.8
```

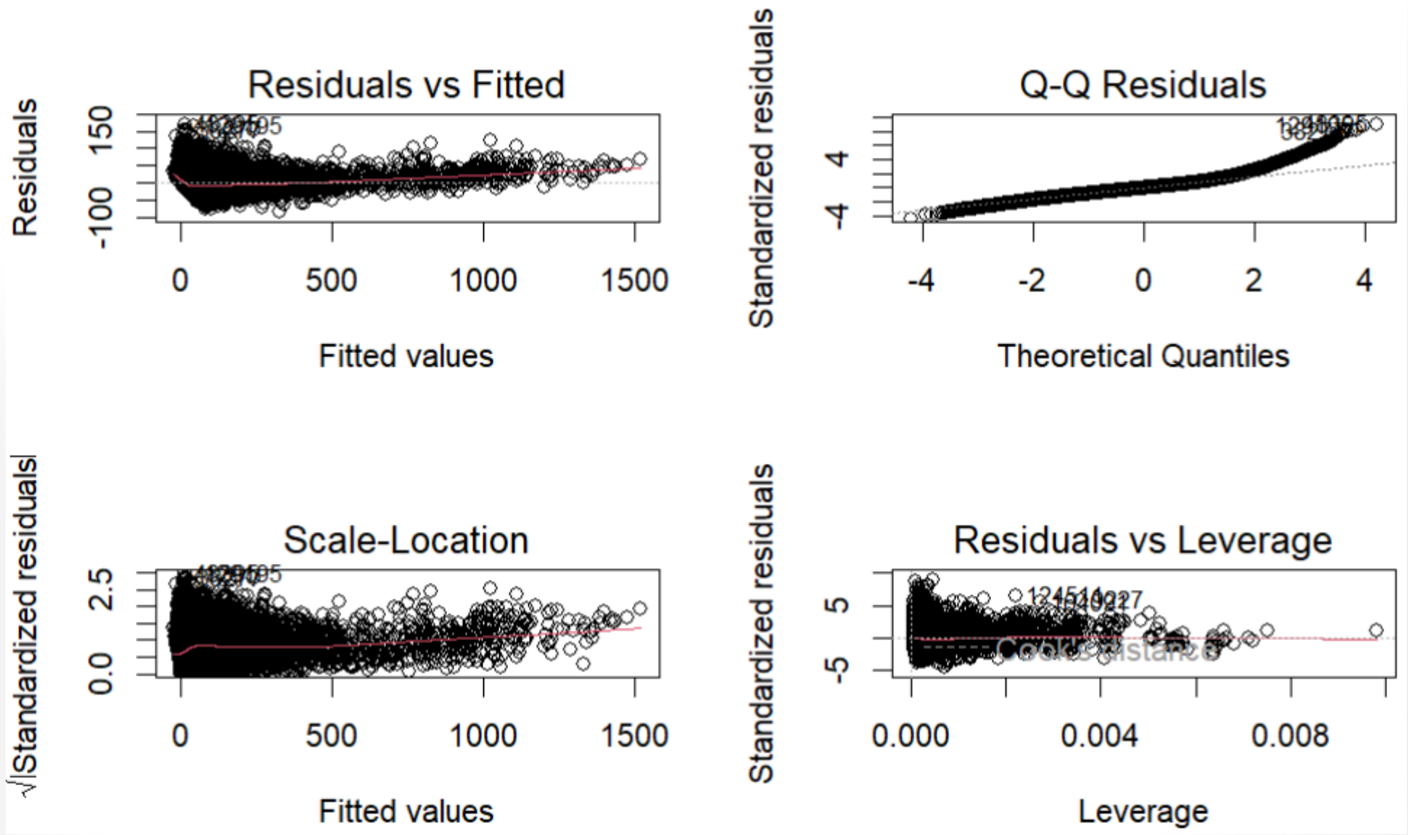
Backward stepwise procedure #2




```
      tavg
     FALSE
```

```
Aicraft_age
       TRUE
```

FINAL MODEL



Summary

```
Residual standard error: 18.94

Adjusted R-squared:  0.969
```

Collinearity

Dep_Delay	1.025292
Flight_Duration	1.049882
Aicraft_age	1.009058
tmin	1.181597
prcp	1.019826
snow	1.146478
DepTime_categoric	1.022690

Conclusions

Regression models

	MSE
linear regression model	284.1507
polynomial (quadratic) regression model	262.4516
polynomial (cubic) regression model	249.0253
multiple regression model	262.8106

Model data

CLASSIFICATION

Steps

- selecting flights that arrived in the JFK airport of NY
- merge of flights and weather datasets
- omit na values
- delete non useful variables
- create a categorical variable for the moments of the day
- Arr_Delay trasformed in categorical variable

Train and test split

Train from October to December
dataset$FlightDate < "2023-09-30"

Number of delayed flights: 49597
Number of not delayed flights: 79945

Logistic regression

- Model
- Reduced model
- ROC plot

Arr_Delay ~ Dep_Delay + Flight_Duration + Aicraft_age + tavg + tmin + tmax + prcp + snow + wdir + wspd + pres + DepTime_categoric

Coefficients initial model

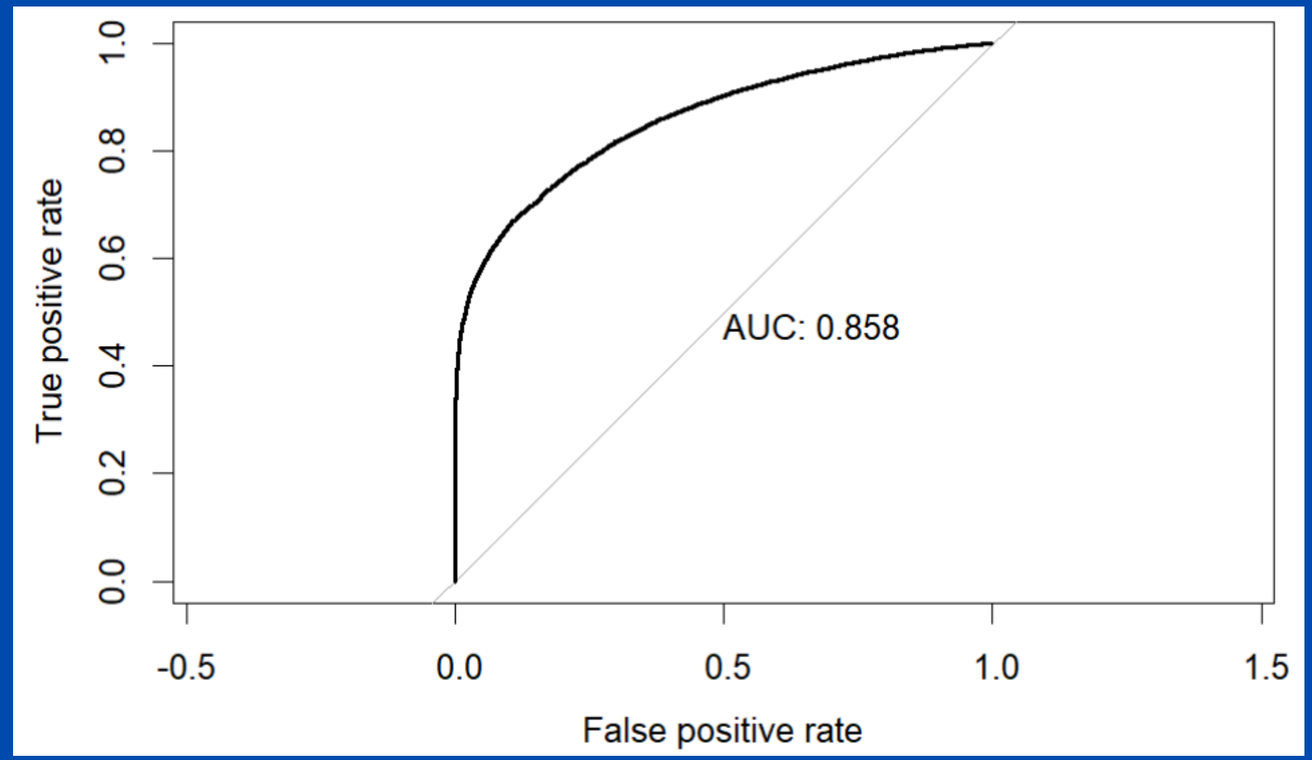
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.4500296	1.7280987	-1.418	0.1563	
Dep_Delay	0.1120685	0.0009600	116.733	< 2e-16	***
Flight_Duration	0.0031344	0.0000963	32.550	< 2e-16	***
Aicraft_age	-0.0050823	0.0011988	-4.239	2.24e-05	***
tavg	0.0078370	0.0092878	0.844	0.3988	
tmin	-0.0313453	0.0052384	-5.984	2.18e-09	***
tmax	0.0022702	0.0050859	0.446	0.6553	
prcp	0.0048327	0.0009696	4.984	6.22e-07	***
snow	0.0023759	0.0003546	6.699	2.09e-11	***
wdir	-0.0003480	0.0000840	-4.143	3.43e-05	***
wspd	0.0039614	0.0016693	2.373	0.0176	*
pres	0.0010045	0.0016864	0.596	0.5514	
DepTime_categoric2	0.4431531	0.0200453	22.108	< 2e-16	***
DepTime_categoric3	0.2409394	0.0236278	10.197	< 2e-16	***
DepTime_categoric4	0.1069042	0.0512978	2.084	0.0372	*

Null deviance: 131918 Residual deviance: 82105 AIC: 82135

Reduced Logistic Regression

(Arr_Delay_Binary ~ Dep_Delay + Flight_Duration + Aicraft_age + tmin + prcp + snow + wdir + DepTime_categoric)

Null deviance: 131918
Residual deviance: 82115
AIC: 82137



LDA

Linear Discriminant Analysis

Arr_Delay ~ Dep_Delay + Flight_Duration + Aicraft_age + tmin + prcp + snow + wdir + DepTime_categoric

Group means:

	Dep_Delay	Flight_Duration	Aicraft_age	tmin
0	-3.162056	172.2426	13.47711	13.22037
1	52.431674	198.7965	13.64494	13.97174

	prcp	snow	wdir
0	2.487985	3.160144	187.1319
1	3.389591	4.030757	188.8948

Coefficients of linear discriminants:

Dep_Delay	1.180273e-02
Flight_Duration	3.187251e-03
Aicraft_age	2.784885e-03
tmin	-1.659446e-04
prcp	9.774186e-03
snow	3.618466e-03
wdir	-4.856806e-05
DepTime_categoric2	7.918853e-01
DepTime_categoric3	7.629473e-01
DepTime_categoric4	-7.925549e-02

Prior probabilities of groups:

0	1
0.5903681	0.4096319

QDA

Quadratic
Discriminant Analysis

Arr_Delay ~ Dep_Delay + Flight_Duration + Aicraft_age + tmin + prcp + snow + wdir + DepTime_categoric

Prior probabilities of groups:

0	1
0.5903681	0.4096319

Group means:

	Dep_Delay	Flight_Duration	Aicraft_age	tmin
0	-3.162056	172.2426	13.47711	13.22037
1	52.431674	198.7965	13.64494	13.97174

	prcp	snow	wdir
0	2.487985	3.160144	187.1319
1	3.389591	4.030757	188.8948

Conclusions

Classification models

	ACCURACY
logistic regression model	0.8371578
linear discriminant analysis model	0.7755191
quadratic discriminant analysis model	0.8305169

Thanks for the attention!

Arianna Morandi & Beatrice Spagnolo & Tomas Guarini