Metody Eksploracji Danych
Projekt Zespołowy

ALIAKSANDR KAROLIK
ŁUKASZ KNIGAWKA

Zakres prac

1. Analiza eksploracyjna danych
   - Zmniejszenie objętości danych
   - Zapoznanie się z cechami
   - Zbadanie zakresów wartości
   - Oczyszczenie danych
   - Przeprowadzenie wizualizacji danych
   - Poszukiwanie trendów okresowych i sezonowości
   - Analiza korelacji cech
   - Postawienie hipotezy badawczej
2. Opracowanie modelu regresji
   - Przygotowanie danych wejściowych
   - Dobór hiperparametru regularyzacji (Lasso/Ridge)
   - Porównanie różnych modeli
   - Zastosowanie boostingu i baggingu

Zbiór danych

Celem projektu jest przeanalizowanie danych z biura U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics które śledzi terminowość lotów krajowych obsługiwanych przez dużych przewoźników lotniczych.

1. *flights.csv*
   - 31 kolumn
   - 5819079 wierszy
2. *airlines.csv*
   - 2 kolumny: kod IATA i nazwa linii
   - 14 różnych linii lotniczych
3. *airports.csv*
   - 7 kolumn
   - 31 lotnisk

Cechy opisujące loty (1)

Nazwa kolumny	Opis
YEAR	Year of the Flight Trip
MONTH	**Month of the Flight Trip**
DAY	Day of the Flight Trip
DAY_OF_WEEK	**Day of week of the Flight Trip**
AIRLINE	**Airline Identifier**
FLIGHT_NUMBER	Flight Identifier
TAIL_NUMBER	Aircraft Identifier
ORIGIN_AIRPORT	**Starting Airport**
DESTINATION_AIRPORT	**Destination Airport**
SCHEDULED_DEPARTURE	Planned Departure Time

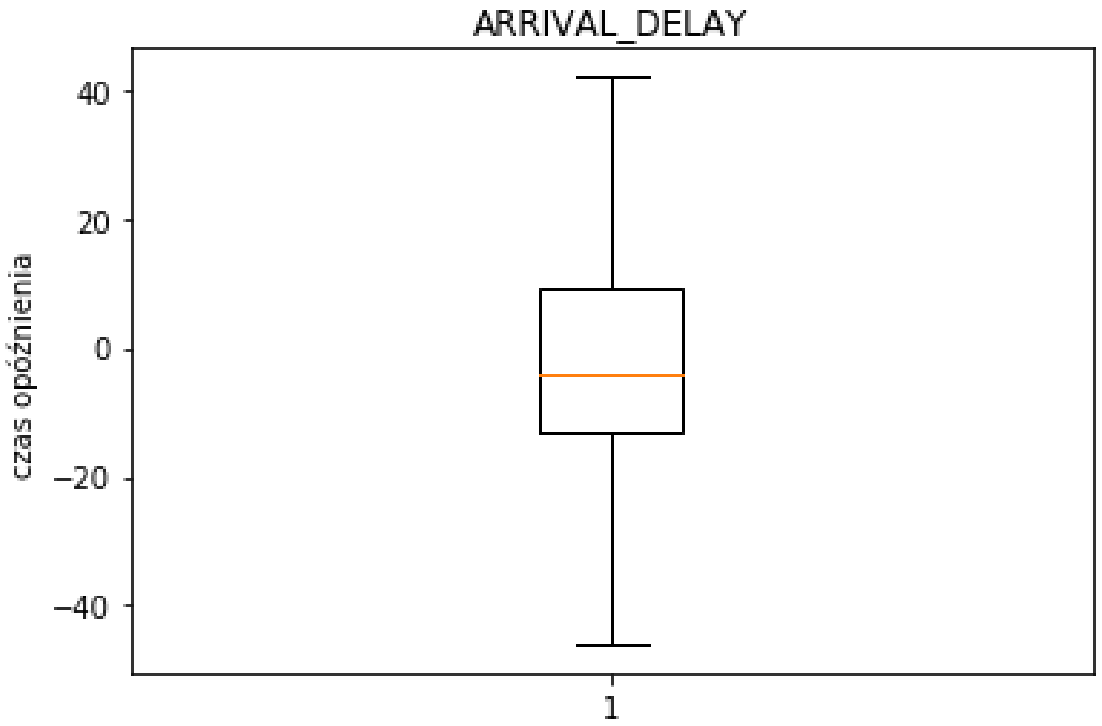
Cechy opisujące loty (2)

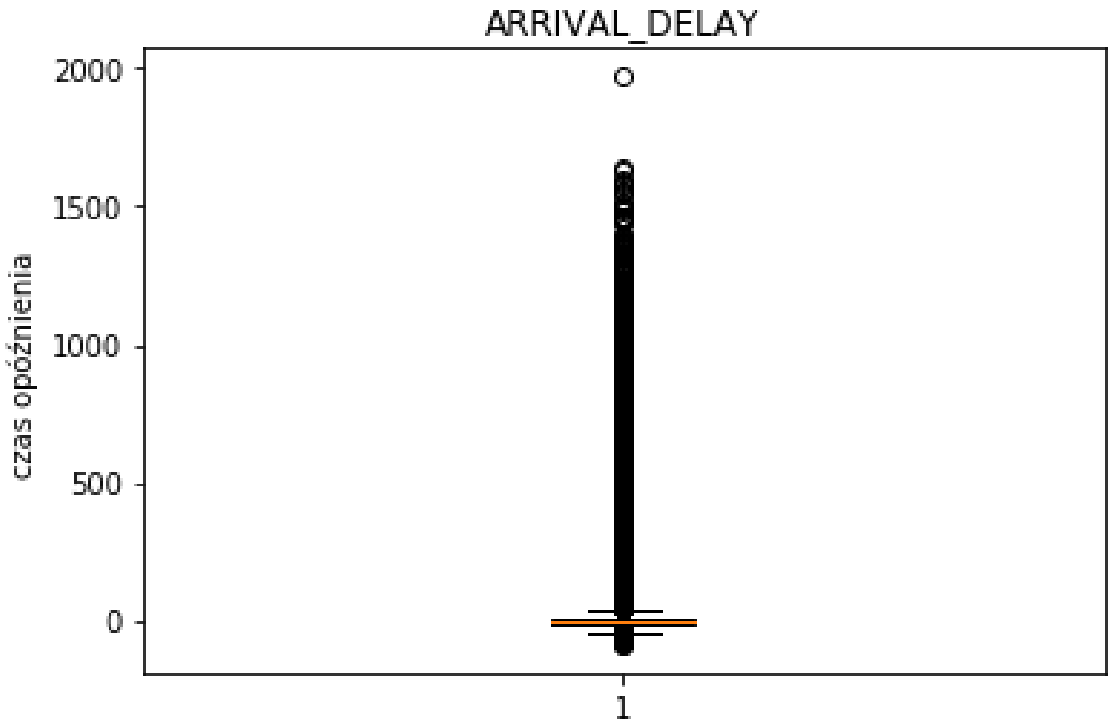
Nazwa kolumny	Opis
DEPARTURE_TIME	TAXI_OUT
DEPARTURE_DELAY	**Total Delay on Departure**
TAXI_OUT	Time between departure airport gate and wheels off
WHEELS_OFF	The time point that the aircraft's wheels leave the ground
SCHEDULED_TIME	**Planned time amount needed for the flight trip**
ELAPSED_TIME	**AIR_TIME+TAXI_IN+TAXI_OUT**
AIR_TIME	**The time duration between wheels_off and wheels_on time**
DISTANCE	**Distance between two airports**
WHEELS_ON	The time point that the aircraft's wheels touch on the ground
TAXI_IN	The time between wheels-on and gate arrival

Cechy opisujące loty (3)

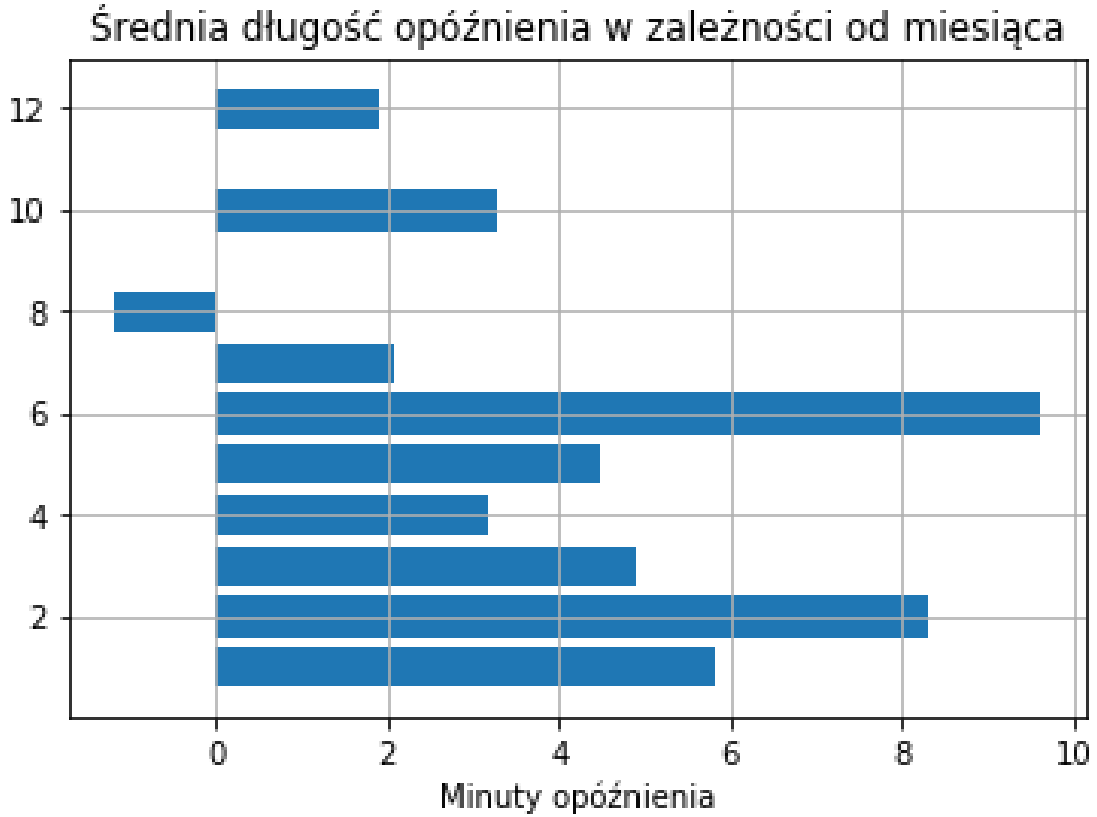
Nazwa kolumny	Opis
SCHEDULED_ARRIVAL	Planned arrival time
ARRIVAL_TIME	WHEELS_ON+TAXI_IN
DIVERTED	Aircraft landed on airport that out of schedule
CANCELLED	Flight Cancelled (1 = cancelled)
CANCELLATION_REASON	Reason for Cancellation of flight
AIR_SYSTEM_DELAY	caused by air system
SECURITY_DELAY	caused by security
AIRLINE_DELAY	caused by the airline
LATE_AIRCRAFT_DELAY	Delay caused by aircraft
WEATHER_DELAY	caused by weather

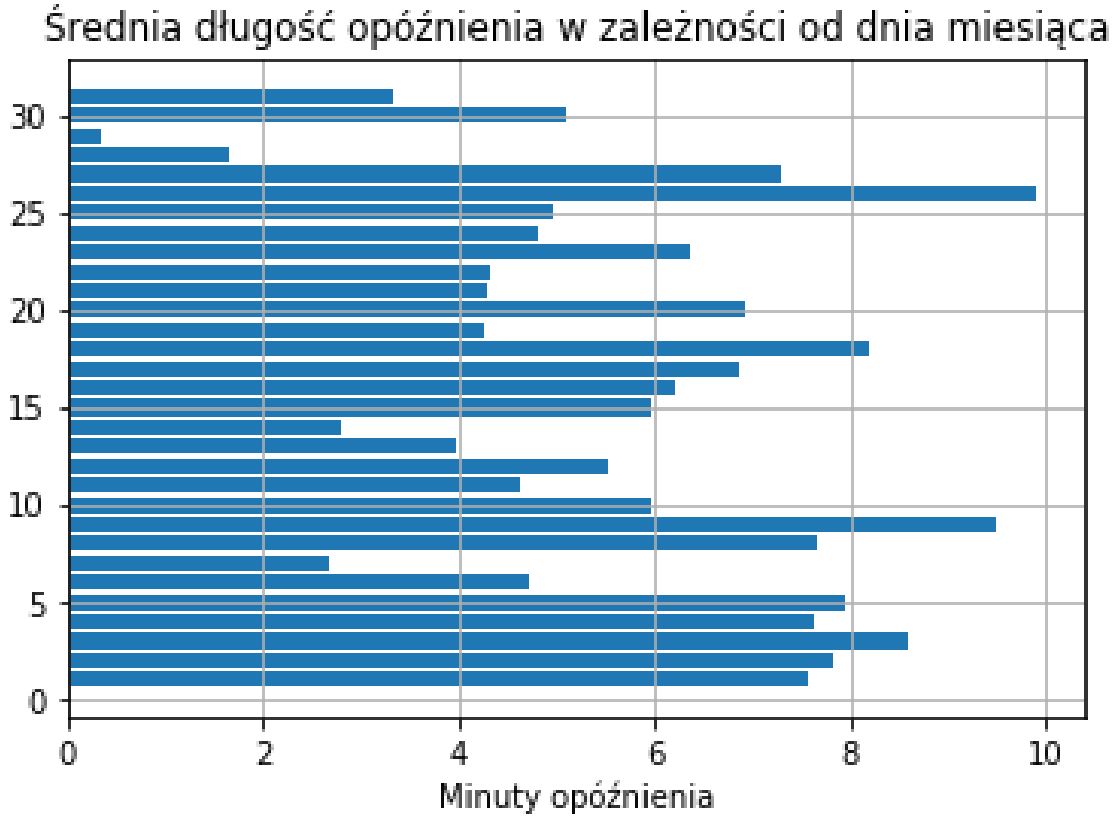
Rozkład cechy opisującej opóźnienie lotu





Długość opóźnienia a miesiąc/dzień miesiąca



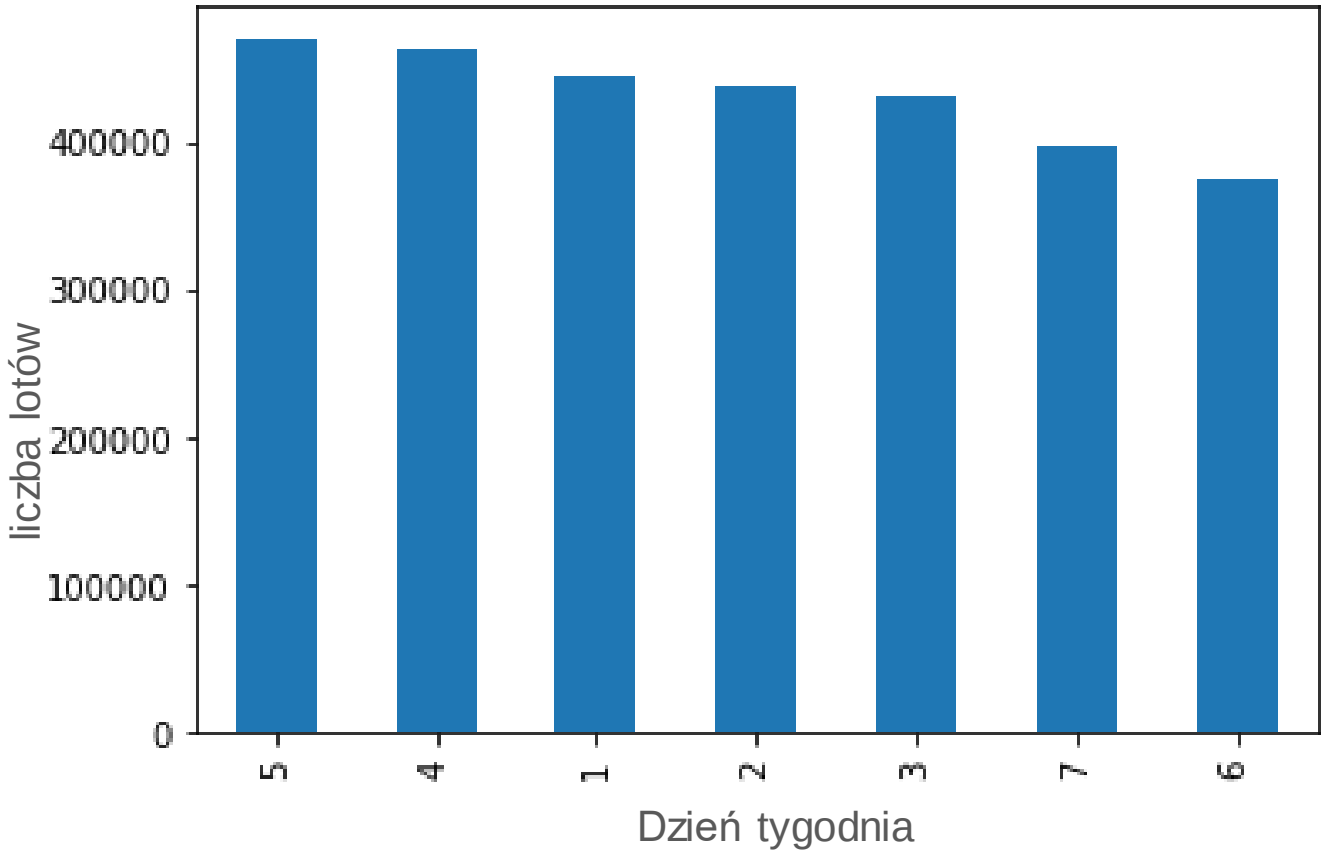


Opóźnienie a dzień tygodnia

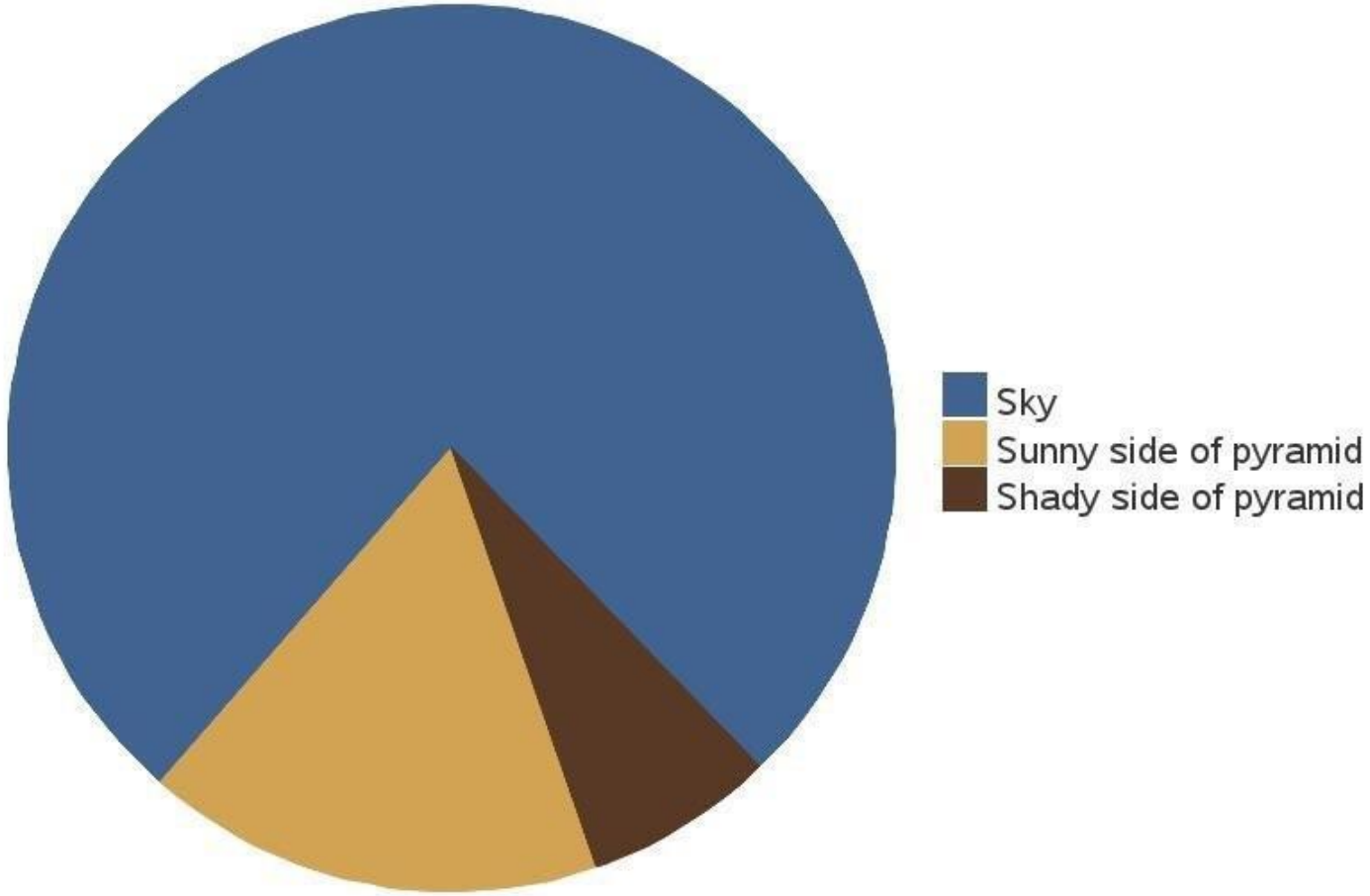
1 - poniedziałek
2 - wtorek
3 - środa
4 - czwartek
5 - piątek
6 - sobota
7 - niedziela

Liczba lotów a dzień tygodnia

1 - poniedziałek
2 - wtorek
3 - środa
4 - czwartek
5 - piątek
6 - sobota
7 - niedziela

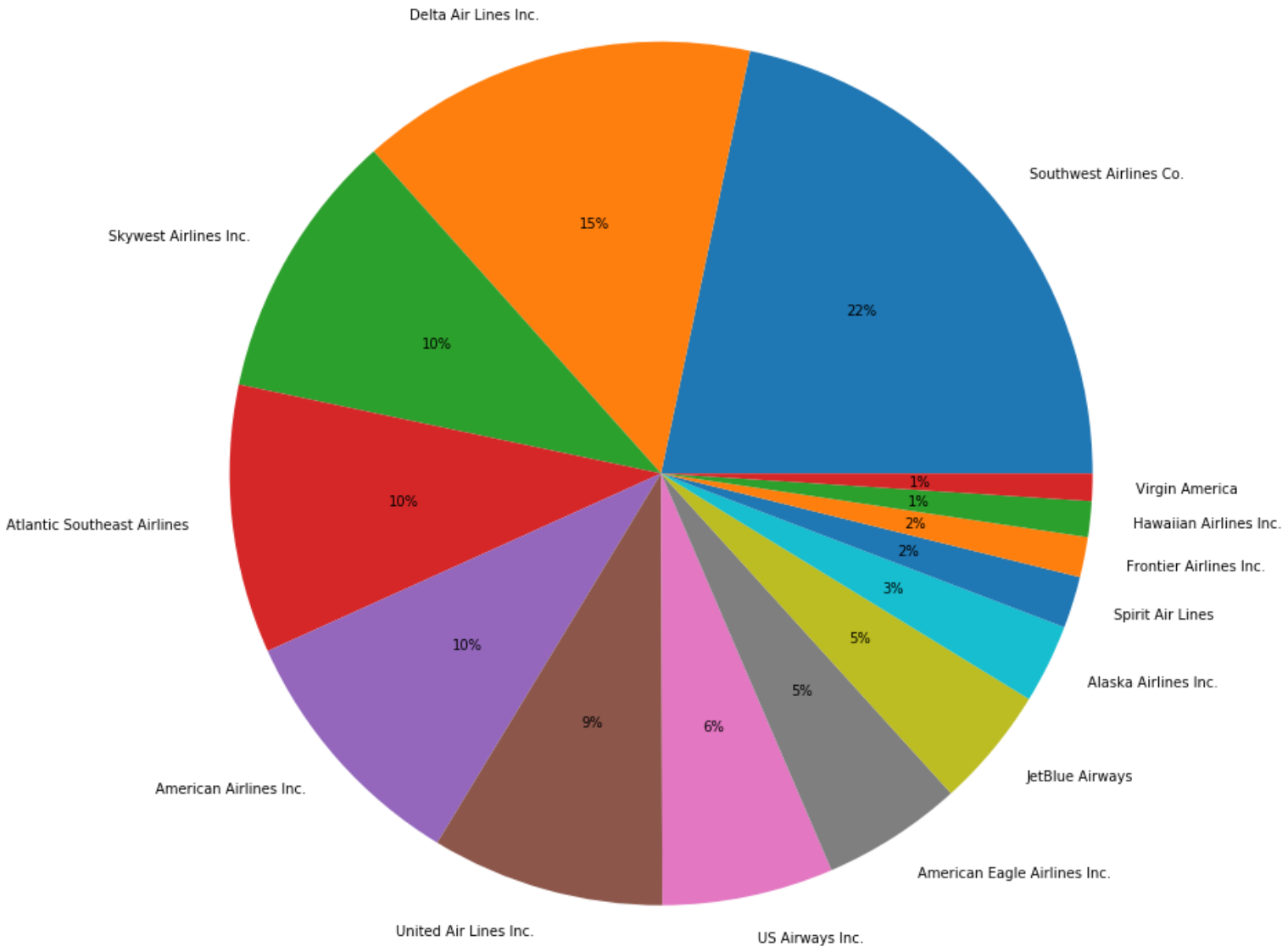


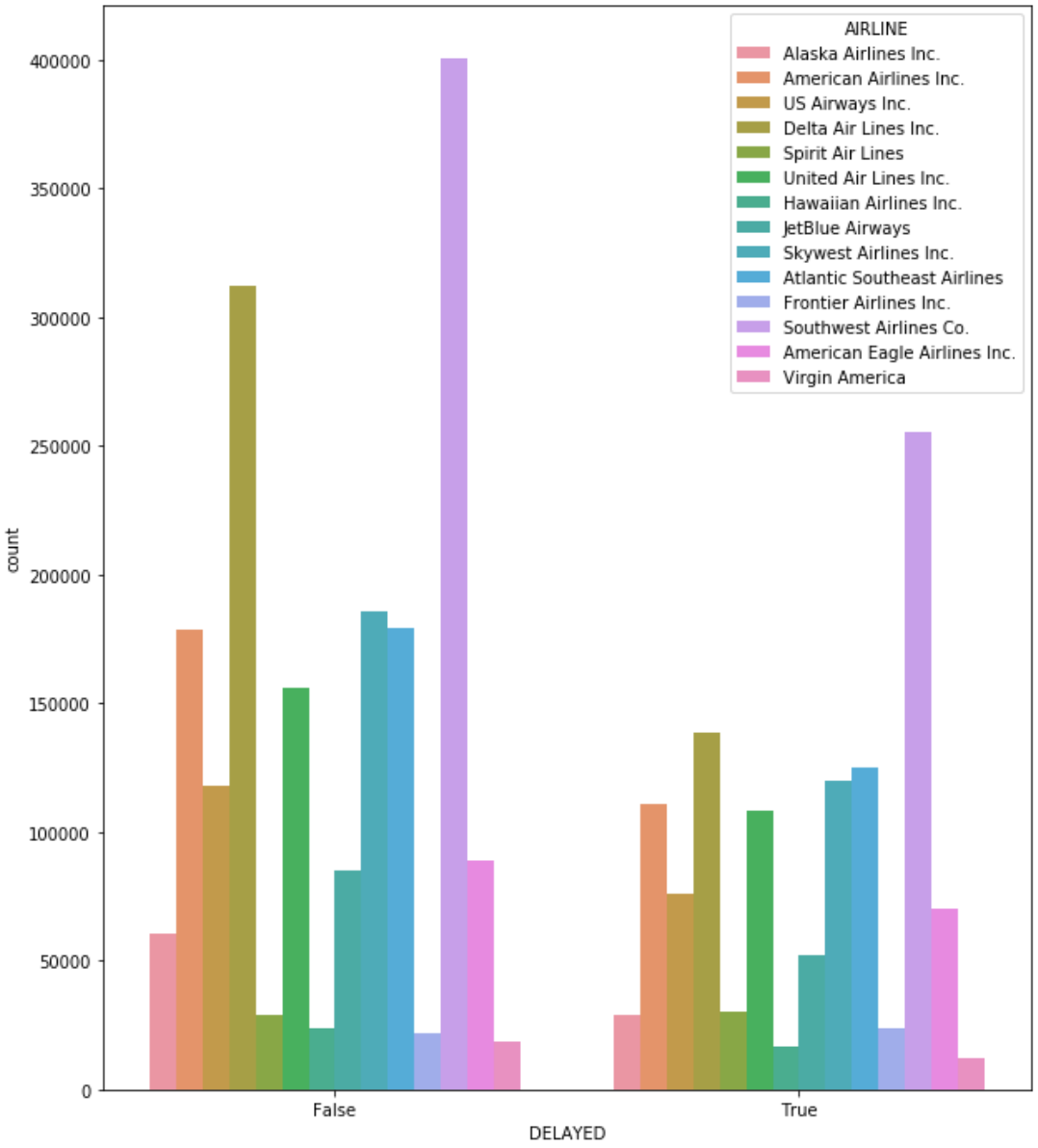
(͡° ͜ʖ ͡°)



Udział linii lotniczych w zbiorze

1. Southwest Airlines (22%)
2. Delta Air Lines (15%)
3. American Airways (12%)
4. Skywest Lines (10%)
5. Atlantic Southeast Lines (10%)





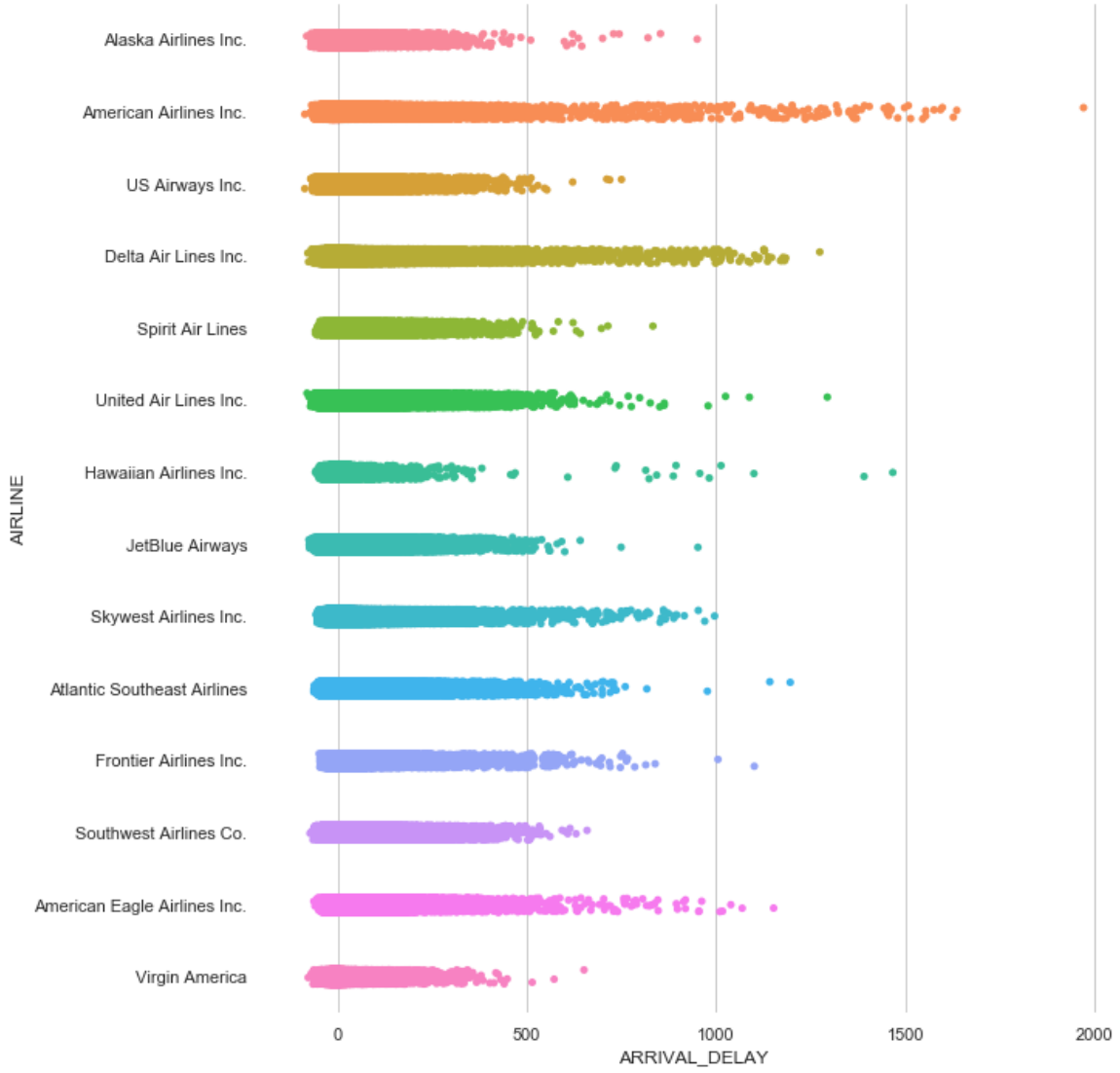
Loty (nie)opóźnione a linia lotnicza

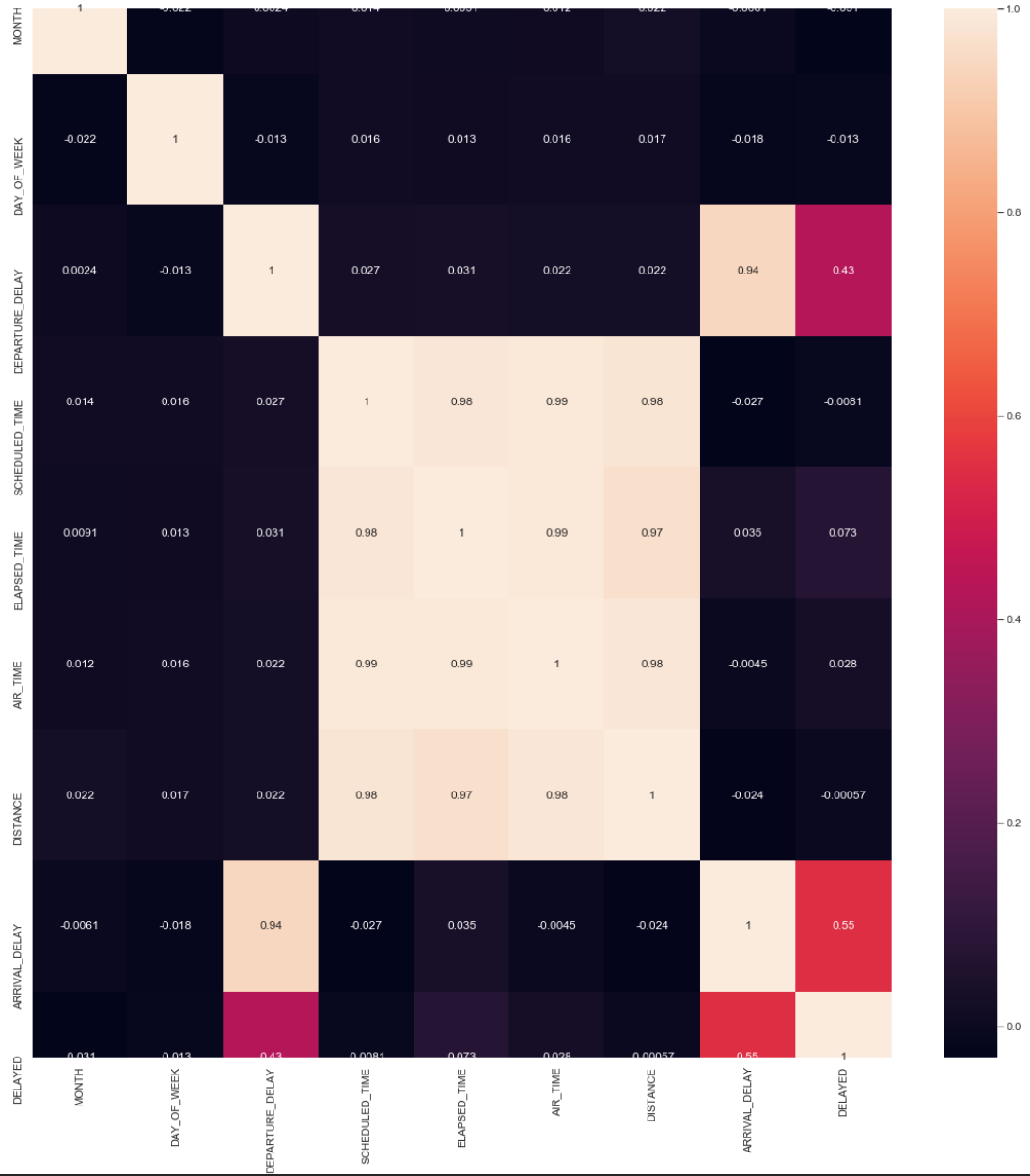
```
flights['DELAYED'] =
flights.loc[:,'ARRIVAL_DELAY'].values > 0
```

Linia lotnicza a wartości opóźnień

American Airlines oraz Delta Air Lines zasługują na naganę.

A może po prostu latają na dłuższych trasach?





Mapa cieplna dla wybranych cech

Zastosowane metody oceny jakości modeli

1. MAE

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

2. MSE == RMSE^2

3. RMSE

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

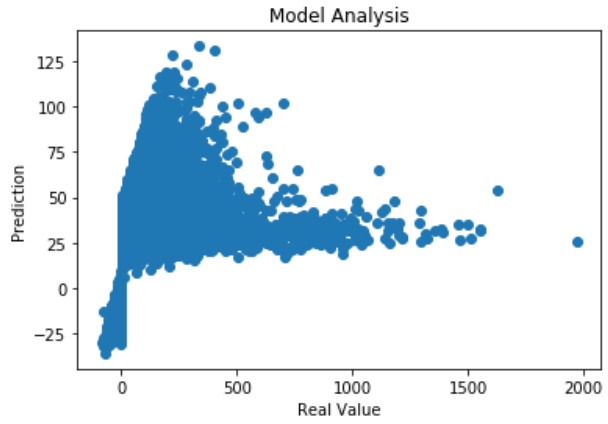
4. R2

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2},$$

Wyniki modeli bez uwzględnienia DEPARTURE_DELAY

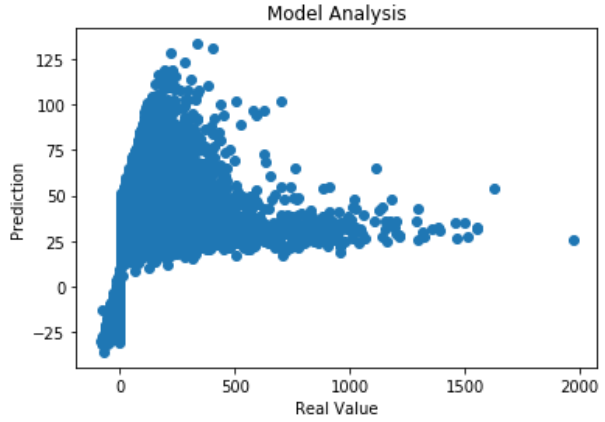
```
Lasso
Mean Absolute Error: 15.283811704880794
Mean Squared Error: 1127.2523119337066
Root Mean Squared Error: 33.57457835824162
R2 :  0.3135032425379226
```

Lasso



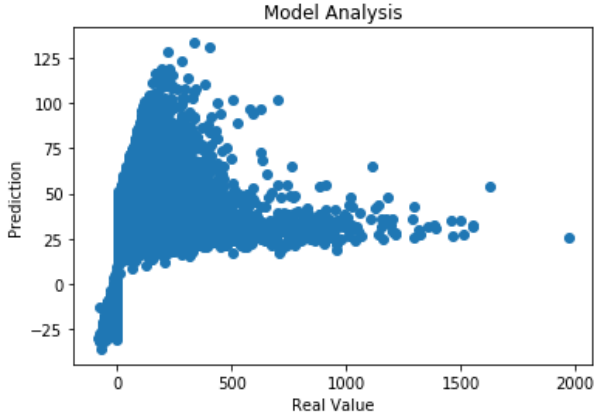

```
Linear
Mean Absolute Error: 15.283811713379887
Mean Squared Error: 1127.2523119569842
Root Mean Squared Error: 33.574578358588276
R2 :  0.3135032425237465
```

Linear




```
Ridge
Mean Absolute Error: 15.283811713380311
Mean Squared Error: 1127.2523119569848
Root Mean Squared Error: 33.57457835858828
R2 :  0.31350324252374606
```

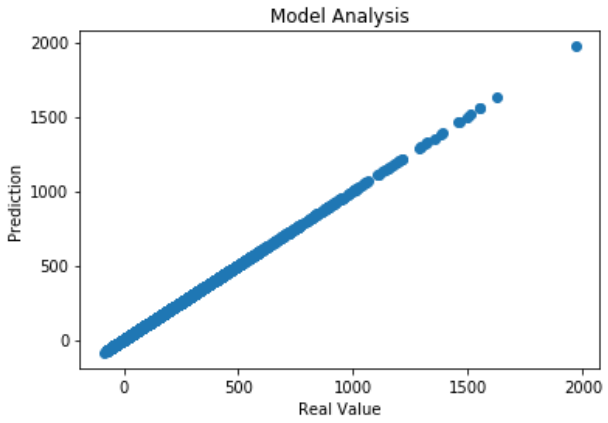
Ridge



Wyniki modeli z uwzględnieniem DEPARTURE_DELAY

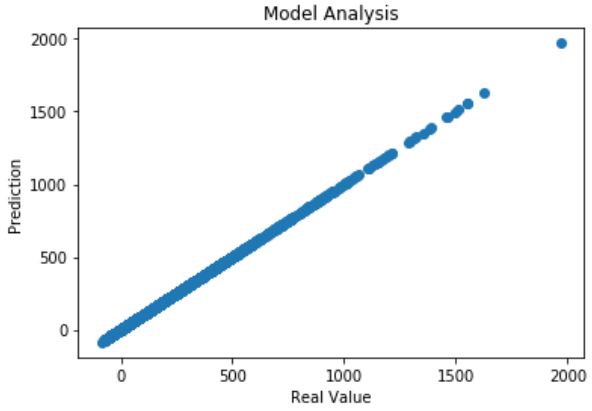
```
Lasso
Mean Absolute Error: 0.038227810334522064
Mean Squared Error: 0.0027469982837444577
Root Mean Squared Error: 0.052411814352724496
R2 :  0.9999983270778028
```

Lasso



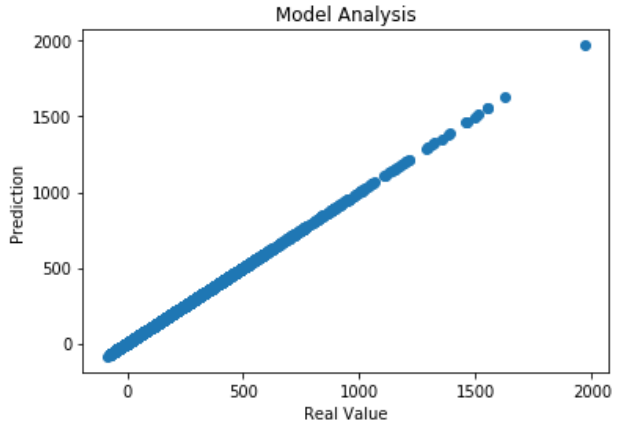

```
Linear
Mean Absolute Error: 5.6354903758030714e-14
Mean Squared Error: 6.908454629085402e-27
Root Mean Squared Error: 8.311711393621294e-14
R2 :  1.0
```

Linear




```
Ridge
Mean Absolute Error: 2.506040217111387e-12
Mean Squared Error: 1.3768398558972216e-23
Root Mean Squared Error: 3.710579275392484e-12
R2 :  1.0
```

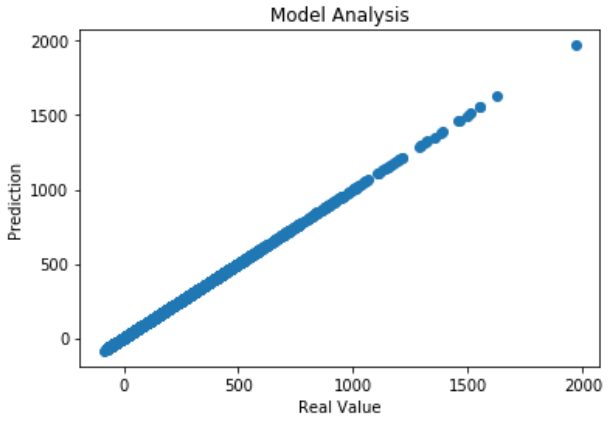
Ridge



Wyniki modeli po zastosowaniu AdaBoost

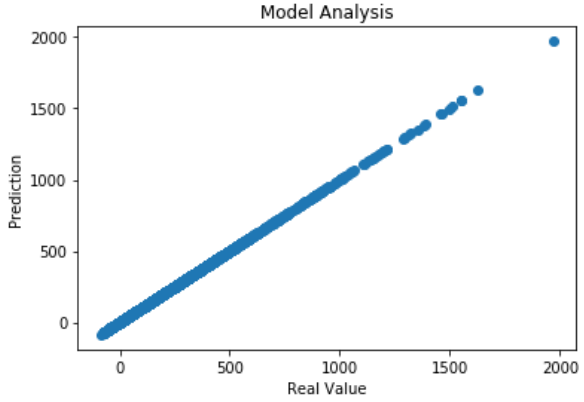
```
Boosted Lasso
Mean Absolute Error: 0.0119537419833688
Mean Squared Error: 0.0002607735880468408
Root Mean Squared Error: 0.016148485627043817
R2 :  0.9999998411888619
```

Boosted Lasso




```
Boosted Linear
Mean Absolute Error: 3.9124343367112725e-14
Mean Squared Error: 2.426653716023703e-27
Root Mean Squared Error: 4.926107708956132e-14
R2 :  1.0
```

Boosted Linear




```
Boosted Ridge
Mean Absolute Error: 4.9134173241012714e-12
Mean Squared Error: 3.978602705309946e-23
Root Mean Squared Error: 6.307616590527634e-12
R2 :  1.0
```

Boosted Ridge