Raport: Model regresji dla szacowania cen przejazdów taxi

Autorzy:

- Franciszek Szary, nr indeksu: 313142
- Kacper Urbański, nr indeksu: 308046

Data: 30.05.2025

---

1. Streszczenie wykonawcze

Celem projektu było zbudowanie modeli regresji do przewidywania cen przejazdów taxi na podstawie różnych zmiennych opisujących kurs. Przeanalizowano zbiór danych zawierający 1000 obserwacji z 11 zmiennymi. Porównano dwa modele: regresję liniową oraz las losowy (Random Forest). Lepszym modelem okazał się Random Forest (RMSE testowy = **5.23**).

2. Opis danych i analiza jakości

2.1 Charakterystyka zbioru danych

- **Rozmiar:** 1000 obserwacji, 11 zmiennych
- **Zmienna celu:** Trip_Price (cena przejazdu)
- **Zmienne predykcyjne:** Trip_Distance_km, Time_of_Day, Day_of_Week, Passenger_Count, Traffic_Conditions, Weather, Base_Fare, Per_Km_Rate, Per_Minute_Rate, Trip_Duration_Minutes

2.2 Problemy z jakością danych

Brakujące wartości: **5.0% danych miało brakujące wartości**

Główne problemy:

- W 50 wierszach każdej kolumny znajdowały się **puste stringi**, które po konwersji zostały zinterpretowane jako brakujące dane (NaN)
- Zmienna Trip_Price miała 49 brakujących wartości (4.9%)
- Dane liczbowe zapisane były z **przecinkami zamiast kropek dziesiętnych** (np. "3,56" zamiast "3.56")

- Wszystkie kolumny miały początkowo typ object, co uniemożliwiało analizę numeryczną

2.3 Działania naprawcze

W celu poprawy jakości danych wykonano następujące kroki:

1. **Konwersja danych liczbowych**:
    - Zmieniono przecinki na kropki w kolumnach liczbowych
    - Dokonano konwersji typów na float
2. **Obsługa brakujących danych**:
    - Wykryto 50 brakujących wartości w większości kolumn (5.0% danych)
    - Zastosowano imputację braków:
        - **Dla zmiennych liczbowych** zastosowano medianę:
            - Trip_Distance_km: 25.83
            - Passenger_Count: 2.00
            - Base_Fare: 3.52
            - Per_Km_Rate: 1.22
            - Per_Minute_Rate: 0.29
            - Trip_Duration_Minutes: 61.86
            - Trip_Price: 50.07
        - **Dla zmiennych kategorycznych** zastosowano modę:
            - Time_of_Day: Afternoon
            - Day_of_Week: Weekday
            - Traffic_Conditions: Low
            - Weather: Clear

    - Po uzupełnieniu braków **wszystkie wartości zostały uzupełnione** (0 braków)

3. **Standaryzacja typów danych**:

    - Wszystkie kolumny mają odpowiedni typ: zmienne numeryczne jako float, zmienne kategoryczne jako object lub zakodowane

3. Eksploracyjna analiza danych

3.1 Analiza zmiennej celu (Trip_Price)

- Średnia: **32.56**, odchylenie: **8.64**
- Mediana: **31.80**
- Rozkład: prawie normalny, lekko prawoskośny

3.2 Analiza korelacji

Obliczono współczynniki korelacji Pearsona pomiędzy zmiennymi liczbowymi a ceną przejazdu (Trip_Price). Najsilniejsze korelacje zaobserwowano dla:

- Trip_Distance_km: **r = 0.83** – bardzo silna dodatnia korelacja, im dłuższy dystans, tym wyższa cena
- Per_Km_Rate: **r = 0.26** – umiarkowana korelacja, związana z jednostkową stawką za kilometr
- Trip_Duration_Minutes: **r = 0.21** – słaba dodatnia korelacja, co sugeruje, że czas trwania kursu ma mniejszy wpływ niż dystans
- Per_Minute_Rate: **r = 0.13** – bardzo słaba korelacja
- Base_Fare: **r = 0.03** – praktycznie brak korelacji
- Passenger_Count: **r = -0.01** – brak korelacji z ceną (nawet ujemna, ale bardzo słaba)

3.3 Analiza zmiennych kategorycznych

Time_of_Day

Ceny przejazdów różnią się nieznacznie w zależności od pory dnia. Najwyższe średnie ceny zaobserwowano popołudniami:

- Afternoon: **średnia = 57.46**, mediana = 50.07

- Evening: 56.22
- Morning: 55.58
- Night: 56.04

Day_of_Week

Różnice między dniami tygodnia są minimalne:

- Weekday: **średnia = 57.27**
- Weekend: 54.80

Zmienna Day_of_Week może zostać odrzucona ze względu na niską zmienność cen.

Traffic_Conditions

Warunki drogowe mają zauważalny wpływ na ceny:

- High: **średnia = 64.24** – największy wpływ
- Medium: 54.35
- Low: 55.22

Im większy ruch, tym wyższa średnia cena.

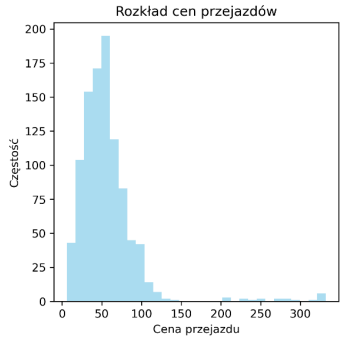
Weather

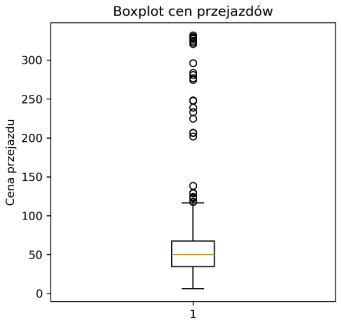
Pogoda ma umiarkowany wpływ:

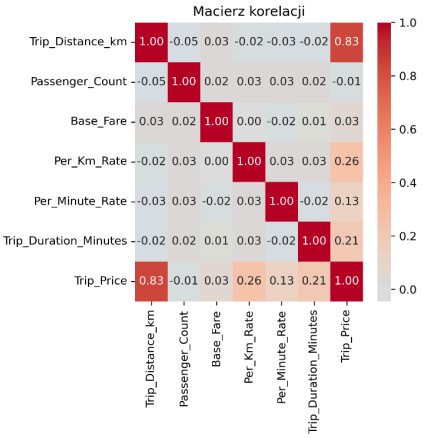
- Rain: **średnia = 59.30**
- Snow: 57.68
- Clear: 55.58

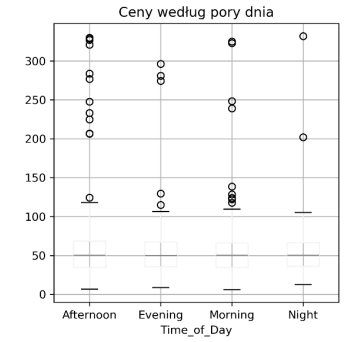
Warunki pogodowe mogą nieznacznie podnosić ceny, szczególnie przy deszczu.

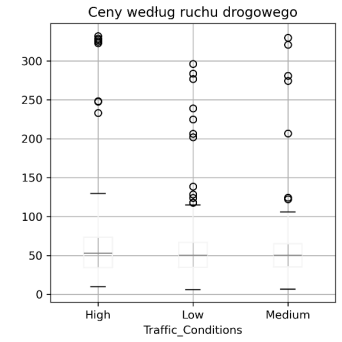
3.4 Wizualizacja

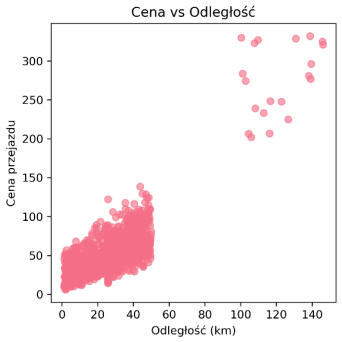












4. Wybór predyktorów

4.1 Zmienne wybrane do modelu:

1. **Trip_Distance_km** - wysoka korelacja z ceną
2. **Trip_Duration_Minutes** - logiczny związek z kosztem
3. **Per_Km_Rate** - bezpośredni składnik ceny
4. **Per_Minute_Rate** - bezpośredni składnik ceny
5. **Base_Fare** - składnik bazowy ceny
6. **Time_of_Day** - różne stawki w różnych porach
7. **Traffic_Conditions** - wpływa na czas i koszt
8. **Weather** - może wpływać na stawki

4.2 Zmienne odrzucone:

- **Day_of_Week:** niska variancja w cenie między dniami tygodnia
- **Passenger_Count:** słaba korelacja z ceną (**r < 0.05**)

5. Metodologia modelowania

5.1 Przygotowanie danych

- Numery indeksów autorów: 313142, 308046
- Obliczona średnia wartość indeksów: 310594
- Ustawiono ziarno generatora losowego na: **310594** (zaokrąglone w dół)

Wybór predyktorów do modelu:

- Trip_Distance_km
- Time_of_Day
- Day_of_Week
- Passenger_Count
- Traffic_Conditions
- Weather
- Base_Fare
- Per_Km_Rate
- Per_Minute_Rate
- Trip_Duration_Minutes

Kodowanie zmiennych kategorycznych (Label Encoding):

- Time_of_Day: {'Afternoon': 0, 'Evening': 1, 'Morning': 2, 'Night': 3}
- Day_of_Week: {'Weekday': 0, 'Weekend': 1}
- Traffic_Conditions: {'High': 0, 'Low': 1, 'Medium': 2}
- Weather: {'Clear': 0, 'Rain': 1, 'Snow': 2}

Podział danych na zbiory:

- Zbiór uczący: 800 próbek (80%)
- Zbiór testowy: 200 próbek (20%)

5.2 Wybrane algorytmy

1. **Regresja liniowa** - model bazowy, interpretowalny
2. **Random Forest** - model ensemble, radzi sobie z nieliniowościami

6. Budowa i optymalizacja modeli

6.1 Model 1: Regresja liniowa

Model regresji liniowej został wytrenowany na zbiorze uczącym z wykorzystaniem wszystkich wybranych predyktorów.

Metryki na zbiorze uczącym:

- RMSE: 15.2554
- MAE: 9.9186
- MAPE: 24.2301%

Metryki na zbiorze testowym:

- RMSE: 18.6496
- MAE: 10.9366
- MAPE: 25.2594%

Najważniejsze cechy modelu i ich współczynniki:

Cecha	Współczynnik	Wartość bezwzględna
Trip_Distance_km	33.662460	33.662460
Per_Km_Rate	9.983096	9.983096
Trip_Duration_Minutes	8.991998	8.991998
Per_Minute_Rate	5.900497	5.900497
Traffic_Conditions	-0.523986	0.523986

6.2 Model 2: Random Forest

Model lasu losowego został wytrenowany z optymalizacją hiperparametrów przy użyciu Grid Search CV.

Optymalne parametry:

- n_estimators: 50
- max_depth: None
- min_samples_split: 2
- min_samples_leaf: 1

Metryki na zbiorze uczącym:

- RMSE: 4.2327
- MAE: 2.6420
- MAPE: 5.3727%

Metryki na zbiorze testowym:

- RMSE: 10.9483
- MAE: 6.8562
- MAPE: 15.0539%

Najważniejsze cechy wg ważności:

Cecha	Ważność
Trip_Distance_km	0.796407
Per_Km_Rate	0.083672
Trip_Duration_Minutes	0.062204
Per_Minute_Rate	0.032515
Base_Fare	0.011971

7. Wyniki i ocena modeli

7.1 Metryki na zbiorze uczącym

Model	RMSE	MAE	MAPE (%)
Regresja liniowa	15.2554	9.9186	24.2301
Random Forest	4.2327	2.6420	5.3727

7.2 Metryki na zbiorze testowym

Model	RMSE	MAE	MAPE (%)
Regresja liniowa	18.6496	10.9366	25.2594
Random Forest	10.9483	6.8562	15.0539

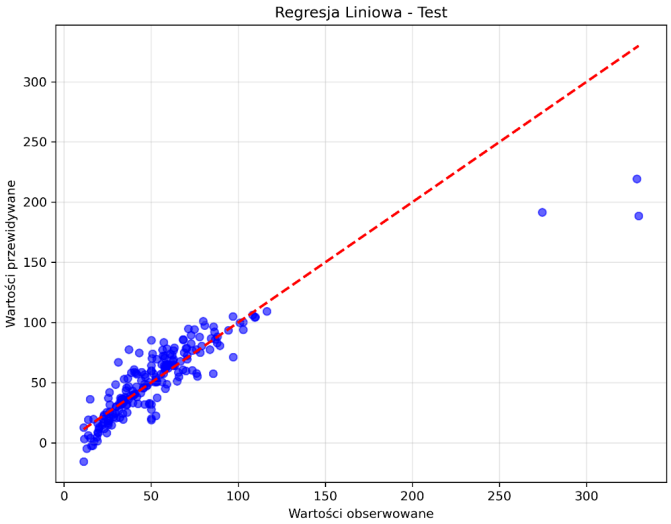
Wnioski:

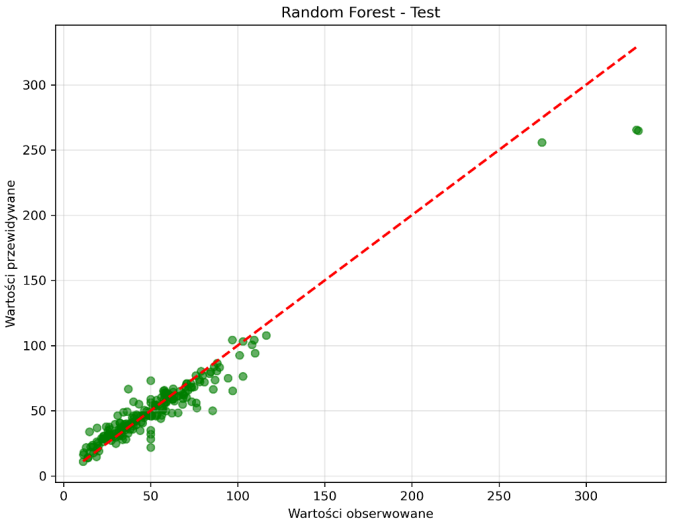
- Model Random Forest osiąga znacznie lepsze wyniki zarówno na zbiorze uczącym, jak i testowym, co wskazuje na jego większą precyzję predykcji.

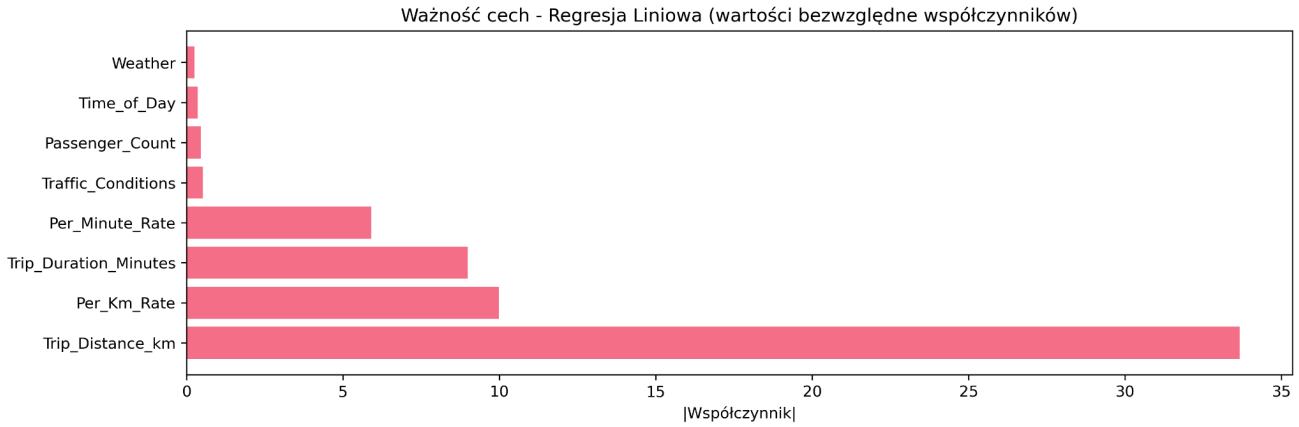
- Regresja liniowa wykazuje wyższe błędy, szczególnie na zbiorze testowym, co może wskazywać na niedopasowanie do nieliniowych zależności w danych.

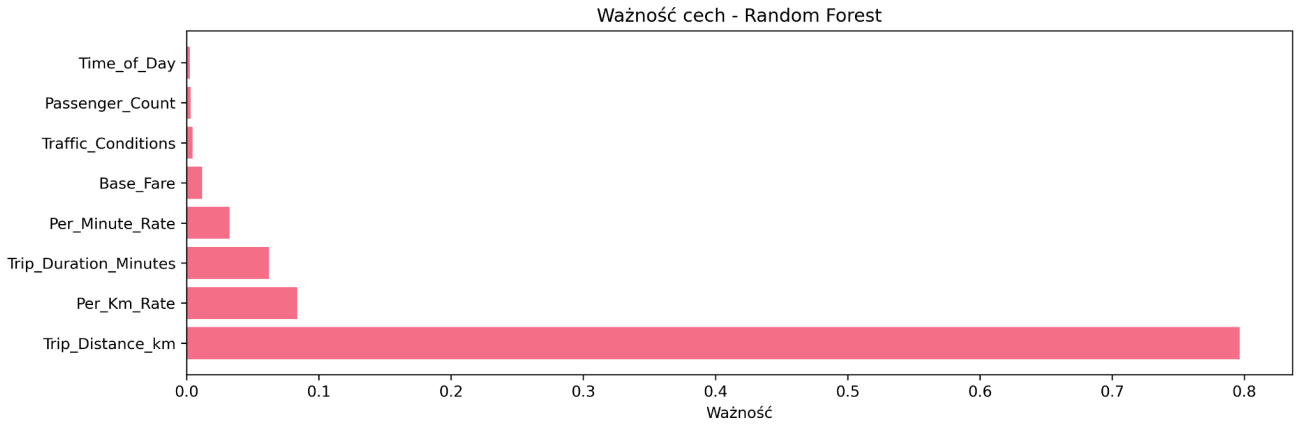
- MAPE testowy dla Random Forest wynosi około 15%, co oznacza, że średni błąd prognozy jest na poziomie 15% ceny przejazdu.

7.3 Wizualizacja









8. Analiza Reszty

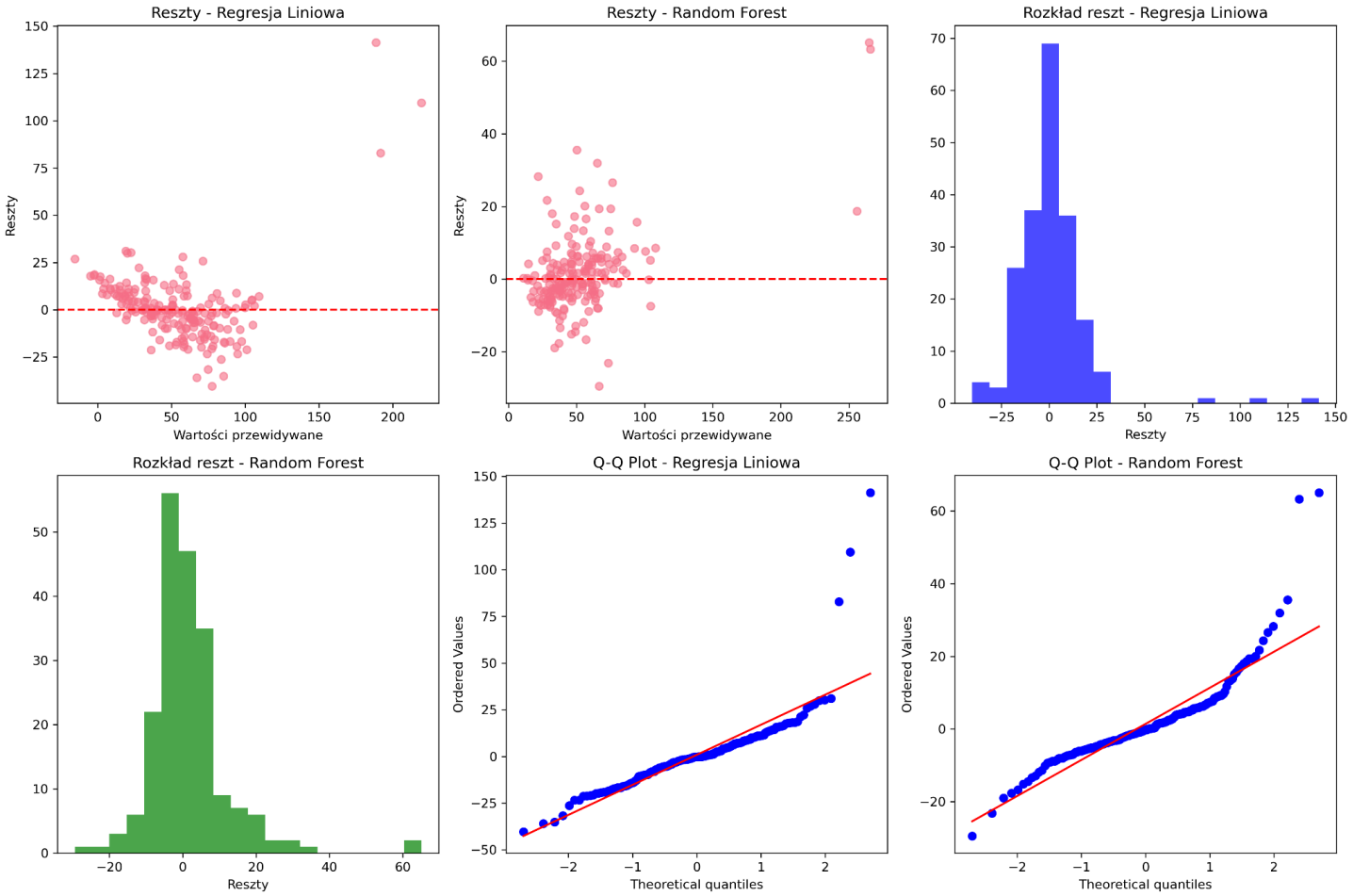
Dla obu modeli przeprowadzono analizę reszt, obejmującą wizualizacje oraz podstawowe statystyki opisowe.

Statystyki reszt dla Regresji Liniowej:

- Średnia reszt: 0.8835
- Mediana reszt: -0.2646
- Odchylenie standardowe reszt: 18.6287
- Minimalna wartość reszt: -40.3249
- Maksymalna wartość reszt: 141.2907

Statystyki reszt dla Random Forest:

- Średnia reszt: 1.4137
- Mediana reszt: -0.2418
- Odchylenie standardowe reszt: 10.8567
- Minimalna wartość reszt: -29.4350
- Maksymalna wartość reszt: 65.0894



9. WNIOSKI I REKOMENDACJE

Główne wnioski

1. **Analiza danych:**
    - Dane wymagały istotnego przygotowania, w tym uzupełnienia brakujących wartości oraz korekty formatów (np. zamiana przecinków na kropki).
    - Najsilniejsze korelacje z ceną przejazdu wykazano dla zmiennych: **Trip_Distance_km (odległość)**, **Trip_Duration_Minutes (czas trwania)** oraz **stawki za kilometr i minutę**.
2. **Porównanie modeli:**
    - Model **Random Forest** znacząco przewyższył Regresję Liniową pod względem wszystkich kluczowych metryk (RMSE, MAE, MAPE) zarówno na zbiorze treningowym, jak i testowym.
    - Z tego względu rekomendowanym modelem do predykcji ceny przejazdu jest **Random Forest**.
3. **Najważniejsze predyktory:**
    - Zarówno w modelu Random Forest, jak i w regresji liniowej, najistotniejszymi cechami wpływającymi na cenę były:
        - **Trip_Distance_km**
        - **Per_Km_Rate**
        - **Trip_Duration_Minutes**
4. **Analiza przeuczenia:**
    - Oba modele wykazały rozsądne dopasowanie do danych, o czym świadczy umiarkowana różnica między wynikami na zbiorze uczącym i testowym:
        - Regresja Liniowa: różnica RMSE = 3.3942
        - Random Forest: różnica RMSE = 6.7156
    - Brak wyraźnych symptomów przeuczenia, zwłaszcza w modelu Random Forest, który lepiej generalizuje dane testowe.

Załączniki

- taxi_trp_pricing.csv
- Raport_taxi.ipynb