الجمهورية العربية السورية
كلية الهندسة المعلوماتية
جامعة دمشق
الاختصاص : الذكاء الصنعي
المادة : استكشاف المعرفة

NEW YORK TAXI CHALLENGE

إعداد الطلاب :

أحمد شلون
إحسان نصير
مؤيد ادريس
ياسر زيدان

اشراف : م .عدنان القطان -علا طبال

2023-6-19

1-عملية معالجة البيانات:

تم تطبيق عمليتان على البيانات قبل البدء باستخدامها وبناء السلسة الزمنية وهما:

a. التخلص من القيم المتطرفة لعمود ال total_amount من اجل استخدامه في السلسة الزمنية.
b. حذف العينات التي تحوي تواريخ خارج مجال (2019-2022)
c. تم تعبئة قيم airport_fee ب اما 1.25 او 0 بناء على id المكان التي انطلقت منه الرحلة.
d. وأخيرا التخلص من الاسطر التي تحوي قيم فارغة

2- بناء السلسلة الزمنية:

تم انشاء السلسة الزمنية حيث قدر عددها ب 67432 حيث كل عينة تحوي ساعة من ساعات النهار وعدد الرحلات وصافي الربح لكل وكالة.

تم التحقق من السلسلة الزمنية ومن استقرارها ثم تعبئة القيم الفارغة للساعات الغير موجودة وحفظها ك long format.

3-هندسة السمات:

- source_zone ,dest_zone ,source_borough ,dest_borough :

تم الانتهاء من هذه السمات عن طريق استخدام python dictionary وذلك لزيادة سرعة التنفيذ بحكم عدد البيانات الهائل حيث يتم تخزين هذا النوع من البيانات من قبل بايثون ك hash table.

- location pair: تم استخدام البنية المذكورة في الأعلى مع استخدام توابع min,max على السلاسل النصية للتخلص من مشكلة التراتبية
- Payment Type, Vendor Name ,RateCode:

تم ارجاع الاسم بناء على ال id الموافق لكل طريقة ولكل اسم.

- Trip Class:

تم حساب عدد الرحلات لكل اسم رحلة عن طريق groupby ومن ثم تم تقسيم الرحلات الى عدة أصناف حيث كل صنف تم تحديده باستخدام تعليمة qcut المضمنة في pandas ومن ثم اسناد الصنف لكل رحلة

- Trip Time:

تمت عن طريق طرح وقت رحلة البداية ورحلة النهاية باستخدام توابع الوقت الموجودة في بايثون لاخذ بعين الاعتبار الفرق في الأيام والشهور.

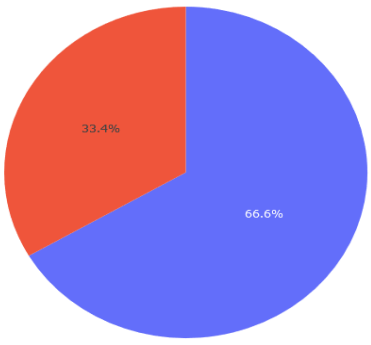
- Trip Distance:

تم حسابها بضرب القيمة السابقة لكل خلية بعمود ال trip_distance ب 1.6 لتحويلها من واحدة الميل الى كيلو متر.

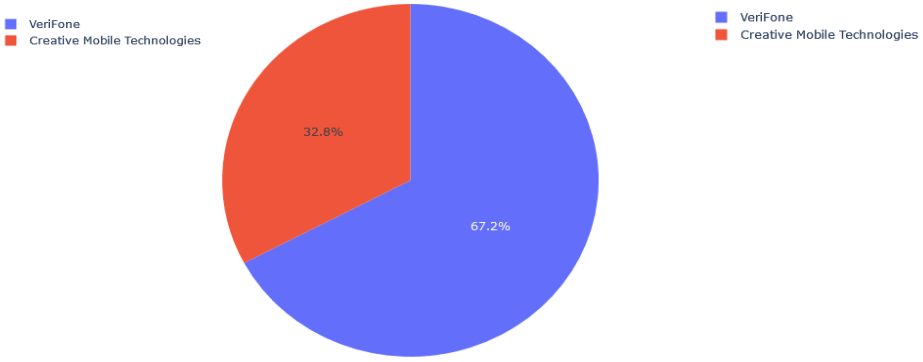
4- الاستكشاف والتحليل:

- حصة كل وكالة:

تم حساب حصة كل وكالة من الأرباح وعدد الرحلات بتجميع البيانات بناء على عمودي ال vendor و ال total_amount وتطبيق عملية الجمع على النتائج.



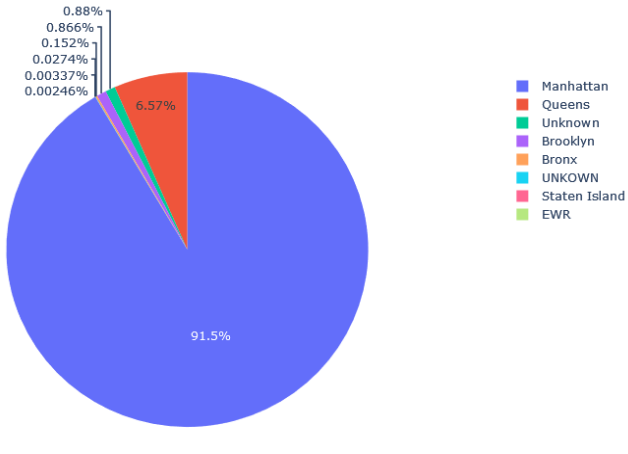
Number Of Journeys

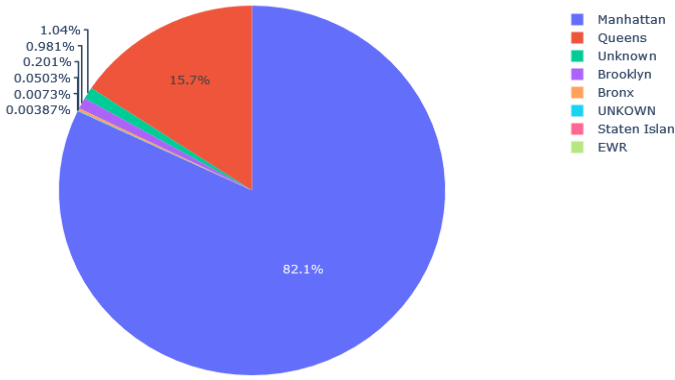


Total Amount

- حصة كل بنديه:

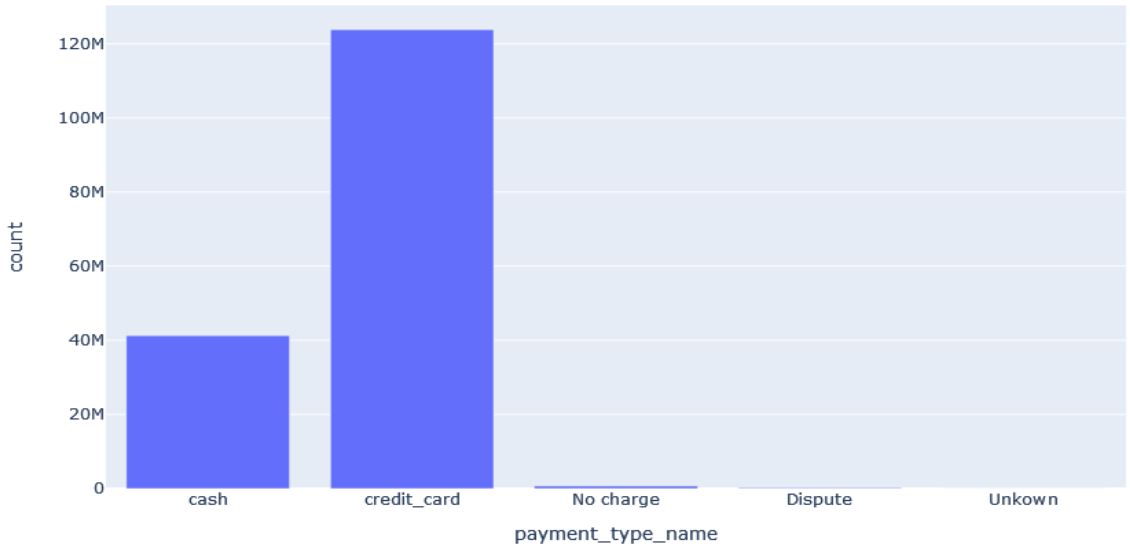
تم حساب أرباح وعدد رحلات كل بلدية بالتجميع على مرة على عمود ال source_borough ومرة على عمود ال distention_borough:





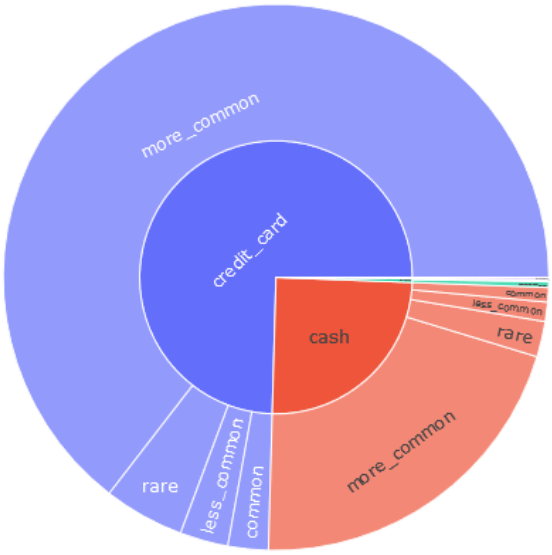
طرق الدفع المستخدمة:

تم الحصول عليها عن طريق التجميع بناء على عمود ال payment_type:



- طرق الدفع المستخدمة لكل صنف:

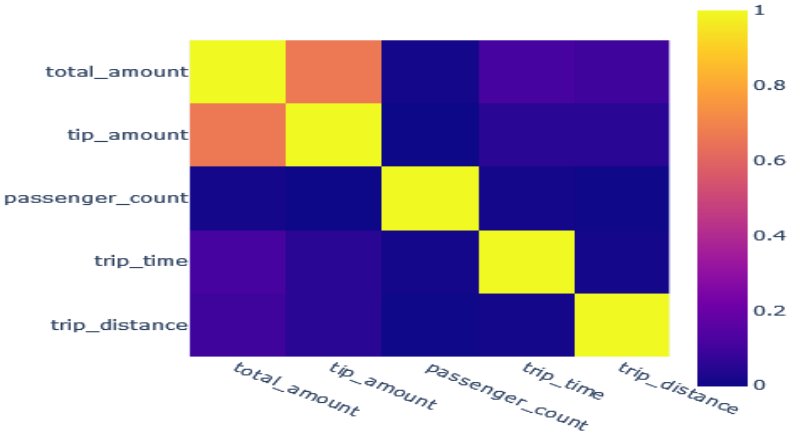
تم حسابها بتجميع البيانات على عمودي ال trip_class و ال pay_type وتطبيق تابع ال size لمعرفة كل رحلة كيف تم الدفع بها:



نلاحظ ان عدد الرحلات التي تم دفعها بالبطاقة اكبر بكثير من باقي الطرق التي تكاد ان تكون معدومة امامها.

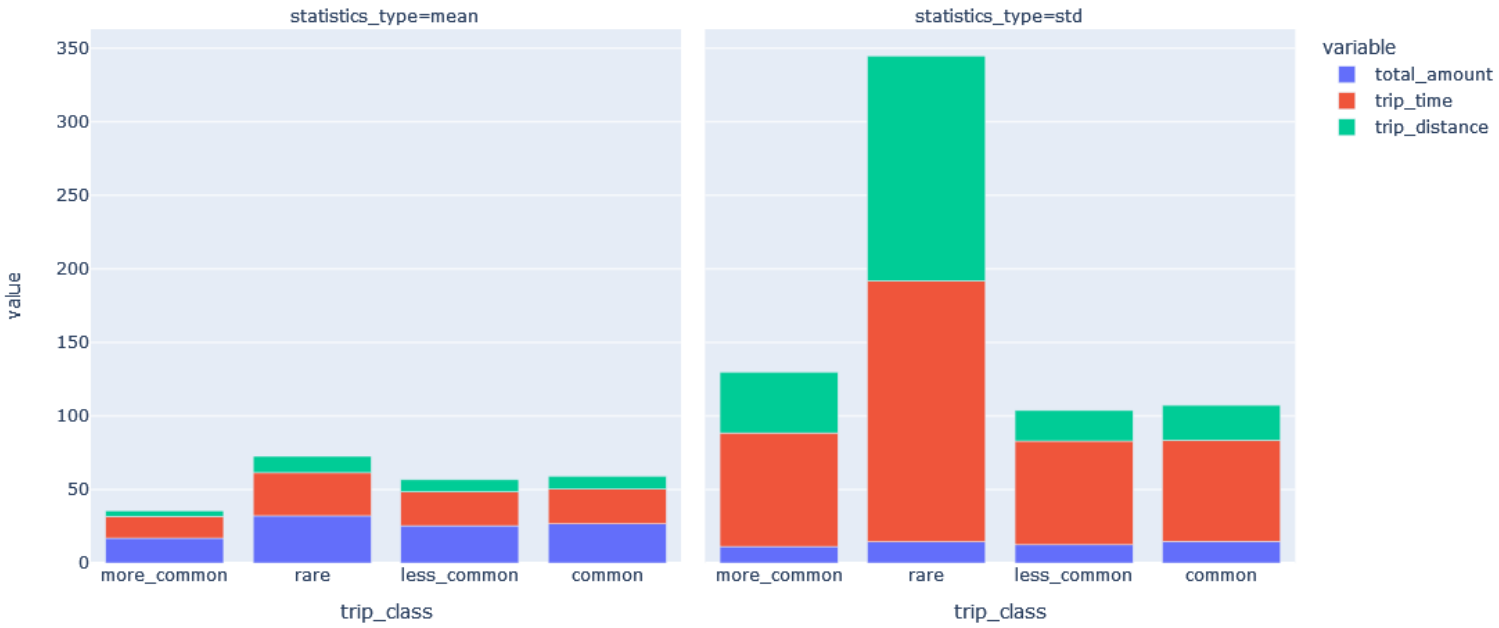
- العلاقة الخطية بين المتغيرات:

لحساب العلاقة الخطية تم حسابها بواسطة بانداس بتالع ال corr ولسهولة القراءة تم عرضها باستخدام heatmap :



حساب المتوسط والانحراف المعيار

تم استخدام التوابع std و mean على الداتا المقسمة بناء على صنف الرحلة وإظهار مجموع الإيرادات ووقت الرحلة و المسافة بالشكل التالي:(trip_time == trip_distance)



- دراسة السلاسل الزمنية:

بعد ان قمنا سابقا بتعبئة الساعات الغير موجودة بالسلسلة الزمنية قمنا بتطبيق عملة downsampling عليها لكي يسهل عرضها وفهمها حيث تم تحويلها الى

1- سلسلة زمنية يومية

2- سلسلة زمنية أسبوعية

3- سلسلة زمنية شهرية

حيث قمنا بدراسة الترابط في السلسلة شهرية واليومية باستخدام توابع ال acf و ال pacf

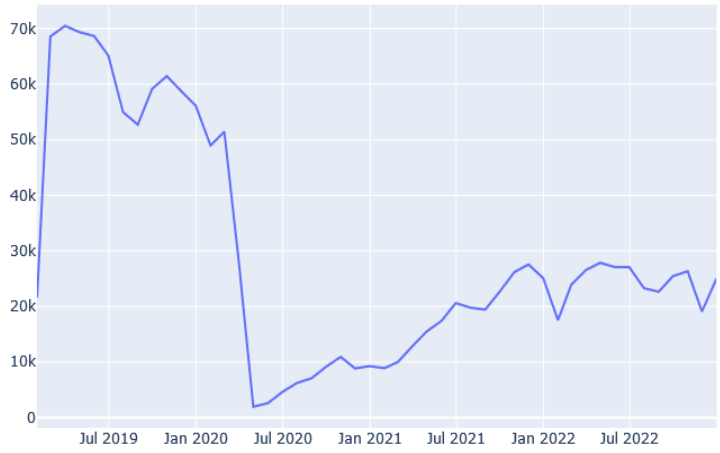
وقمنا باختبار استقرار هاتين السلسلتين عن طريق ال ADF وذلك لمعالجتها قبل إدخالها للموديلات القادمة حيث كانت القيم كالتالي:

الدراسة التالية هي للوكالة الأولى ولكن في الكود تم الدراسة للوكالتين بشكل كامل.

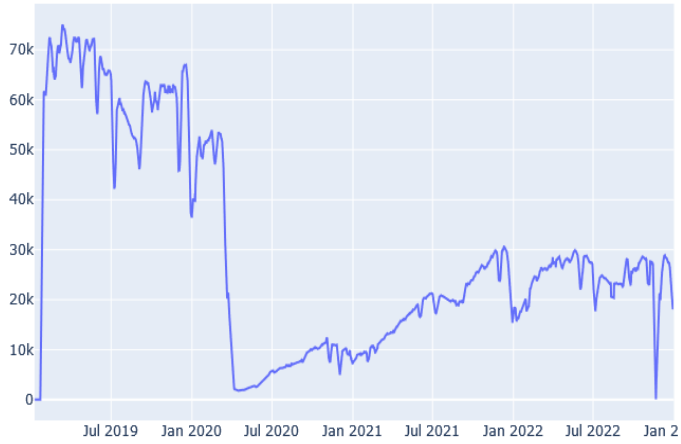
نوع السلسلة	يومية	شهرية
ADF Value	-1.68473	-2.4856
%5	-2.8635	-2.9267
stationarity	NO	NO

لذا ف السلسلة ليست مستقرة ولذلك يجب علينا القيام اما ب differencing او بتطبيق تابع ال log وقمنا بالطريقة الأولى لجعل السلسلة مستقرة.

- من اجل رسم السلسلة الزمنية قمنا أيضا بعمل rolling(168) حوالي الأسبوع تقريبا ورسمها أيضا بالشكل الشهري:



السلسلة الشهرية



السلسلة الاسبوعية

- التدريب والتنبؤ:

تم تقسيم البيانات الى تدريب واختبار تجريب العديد من الموديلات بمختلف التعقيدات وهي:

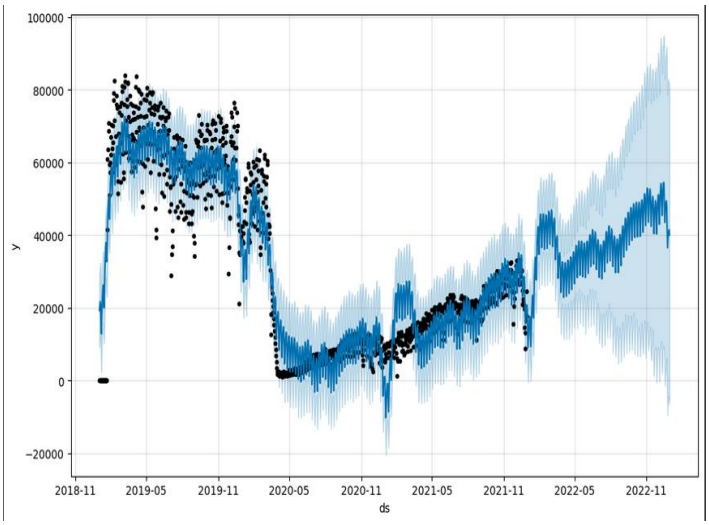
1- Prophet

2- AR with order(3,0,0)

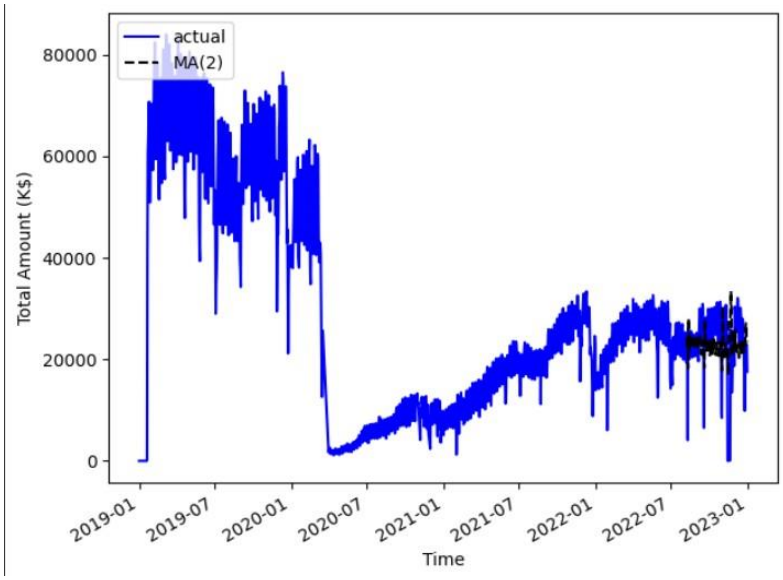
3- ARMA with order(8,0,1)

4- ARMA with order(5,0,1)

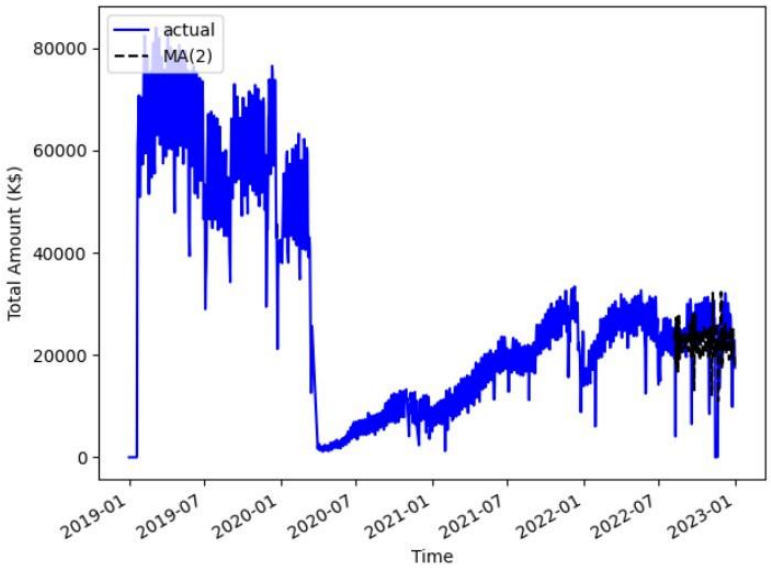
الشكل النهائي للداتا الذي تم التنبؤ به والحقيقي لكل موديل:



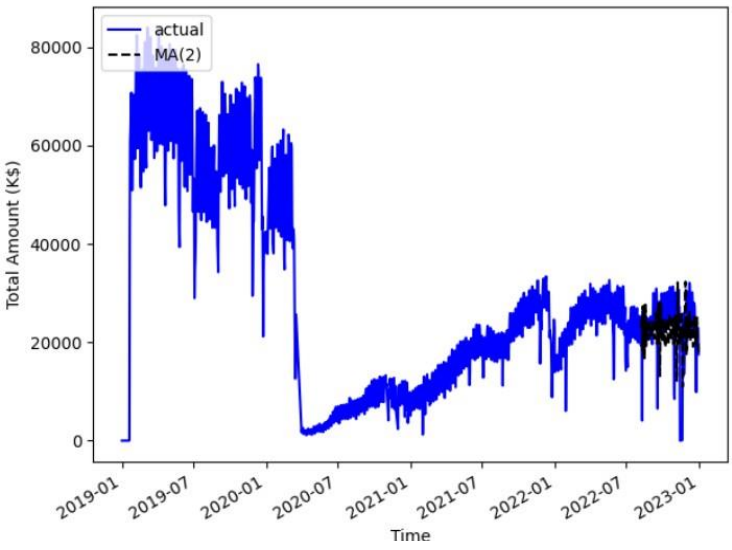
Prophet



AR(3,0,0)



ARMA(8,0,1)



ARMA(5,0,1)