الجمهورية العربية السورية

جامعة دمشق

كلية الهندسة المعلوماتية

قسم الذكاء الصنعي

DM Homework2

تحدي سيارات الأجرة

ساره الغدير

راما ريحاوي

تسنيم عجاج

رغد الحلبي

إشراف

م. عدنان قطان

م. علا طبال

حزيران - 2023

1 تنظيف ومكاملة البيانات

1.1 جوهر الوقت

بعد استكشاف مجموعة البيانات تبين الحاجة إلى عمليات التنظيف التالية:

1.1.1 حذف الأعمدة (airport_fee، mta_tax، improvement_surcharge)

لأن العمود airport_fee معظم قيمه Null، والأعمدة mta_tax
وimprovement_surchargeg معظم قيمها ثابتة.

1.1.2 حذف الأسطر الفارغة

الأعمدة التالية passenger_count، RatecodeID، store_and_fwd_flag،
congestion_surcharge تحتوي قيم فارغة لذا حذفنا الأسطر الفارغة على أساسها.

1.1.3 حذف التعارضات

1.1.3.1 حذف التعارض في التاريخ

- حذف الرحلات المسجلة بتاريخ tpep_pickup_datetime أو
  tpep_dropoff_datetime قبل 2019 أو بعد 2022.
- حذف الرحلات التي تاريخ انطلاقها tpep_pickup_datetime يسبق تاريخ انتهاءها
  tpep_dropoff_datetime.

1.1.3.2 حذف القيم السالبة

حذف الأسطر التي تحوي قيماً سالبة في أحد الأعمدة التالية: extra،
passenger_count، trip_distance، tip_amount، tolls_amount،
congestion_surcharge، total_amount لأن ليس منطقياً أن تحوي قيماً سالبة
(كمية، مسافة، وقت، ...).

1.1.3.3 حذف القيم الخارجة عن التصنيف المحدد

- حذف قيم العمود RatecodeID، عدا القيم المحددة للتصنيف ([1, 2, 3, 4, 5, 6]).
- حذف قيم العمود store_and_fwd_flag، عدا القيمتين (['Y', 'N'])
- حذف قيم العمود payment_type، عدا القيم المحددة للتصنيف ([1, 2, 3, 4, 5, 6])

1.1.4 حذف القيم المتطرفة

تم اكتشاف القيم المتطرفة باستخدام معيار ال z-score الذي يمثل نسبة انحراف القيمة عن الانحراف المعياري.

حيث وجدنا انحرافاً في بعض القيم في الأعمدة التالية:

1.1.4.1 trip_distance

تم حذف القيم التي انحرافها أكبر من 10 لأنها حالات نادرة وربما تؤثر قيمها في الدراسة التحليلية.

1.1.4.2 passenger_count

تم حذف قيمتين بانحراف أكبر من 6 حيث انحرافهما (84 و98) (وهو انحراف كبير بالنسبة لعدد ركاب تكسي أجرة 🙃).

1.1.4.3 fare_amount

تم حذف القيم التي انحرافها أكبر من 100 نظراً لأن معظم القيم انحرافها متقارب وأقل من هذه القيمة عدا بعض القيم الشاذة.

1.1.4.4 Extra

تم حذف القيم التي انحرافاتها أكبر من 100 (اعتبرنا الحد 100 قيمة منطقية مقارنة بباقي القيم ولأن القيمة نقدية)، وهي قيمة وحيدة تحقق الحالة (13191.35).

1.1.5 بناء ال TimeSeries

تقسيم تاريخ الانطلاق إلى سنة، شهر، يوم، ساعة، ثم تجميع الرحلات وفقاً لها ولرقم الوكالة (مع حساب عدد الرحلات، ومجموع الإيرادات الكلي لكل وكالة).

1.2 ضبط آلة الزمن

- إنشاء سمة جديدة datetime ناتجة عن تجميع الأعمدة السابقة وتحويلها إلى datetime.
- إجراء عملية resample على الساعة لكل وكالة لاستكمال الساعات الفارغة في السلسلة.
- تم التحقق من أن الفترات الزمنية بين الرحلات المتتالية في السلسلة الزمنية منتظمة.

2 هندسة السمات

2.1 السمات: source_zone، destination_zone، source_borough، destination_borough

تم إنشاءها نتيجة دمج مجموعة البيانات، مع مجموعة بيانات taxi_zone_lookup.

2.2 location_pair

قمنا بترتيب ال source_zone وال destination_zone لتحقيق خاصة التبادلية للسمة، وجمعهما في سمة واحدة.

2.3 payment_type_name، vendor، rate_code

أجرينا عملية mapping للحصول على تصنيفات كل سمة تبعاً للأعمدة المتعلقة بها وإنشاء السمة الجديدة.

2.4 trip_class

بعد حساب تكرارات ال location_pair، اعتمدنا على حساب الmax وال min وال mean للتكرارات لتحديد مجالات كل صنف. لينتج لدينا التصنيف التالي:

[1, 1143, 69258, 217536, 2129670] = ['rare', 'less-common', 'common', 'more-common']

2.5 trip_duration

حصلنا عليها بطرح وقت الانطلاق من وقت انتهاء الرحلة.

2.6 trip_distance_km

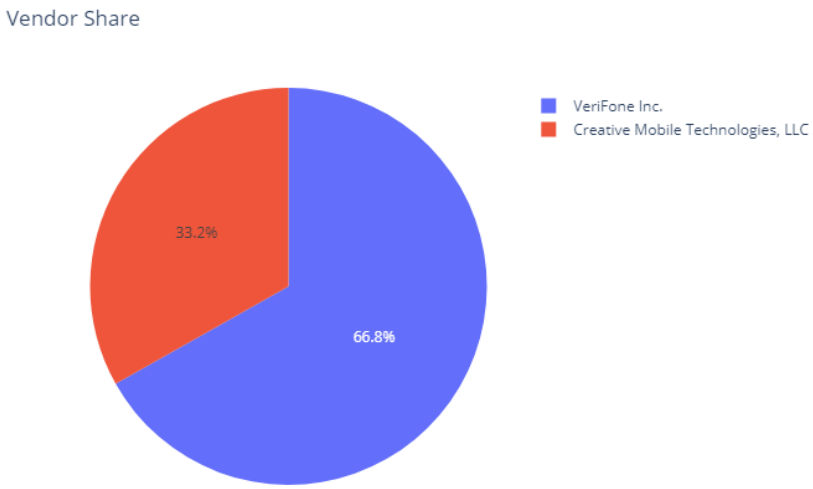
للتحويل من ميل إلى كيلو متر ضربنا ب 1.609344.

3 الاستكشاف والتحليل

3.1 ما وراء حافة الأرقام

3.1.1 حصة كل وكالة

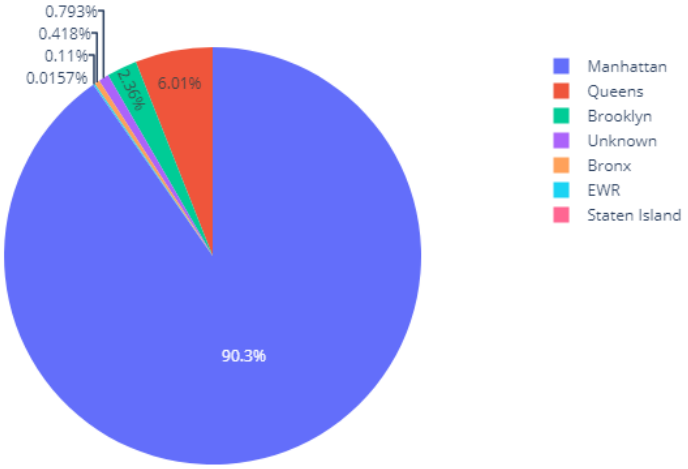
نلاحظ أن حصة الوكالة VeriFone Inc من السوق أكبر من حصة الوكالة الأخرى بمقدار الضعف تقريباً.



بعد إجراء بعض عمليات البحث وجدنا أن وكالة verifone inc هي الأقدم في السوق وقد تأسست بتاريخ 1992 باسم مختلف من قبل سائق تكسي أساساً وقد أصبح اليوم اسمها curb وقد دخلت إلى تكنولوجيا تطبيقات حجز الرحلات لتنافس شركات uber و lyft وأخيراً تعاقدت مع شركة uber في شهر آذار عام 2022 مما قد يكون سببا في سرطتها على السوق مقارنةً بشركة creative mobile technology, llc والتي تأسست في عام 2005 والتي تمتلك العديد من الحلول التكنولوجية مثل تطبيقات الهواتف المحمولة ومواقع الويب وغيرها.

3.1.2 حصة كل بلدية

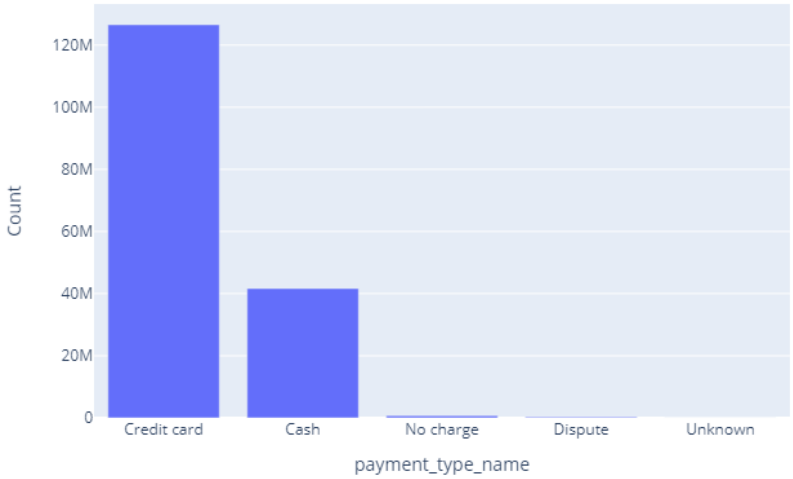
نلاحظ أن بلدية Manhattan تسيطر على السوق.



تُعدّ منهاتن الأكثر شهرة بين جميع البلديات والأكثر ازدحاماً والأكثر زيارةً وقد يكون هذا السبب المحتمل لسيطرتها على السوق.

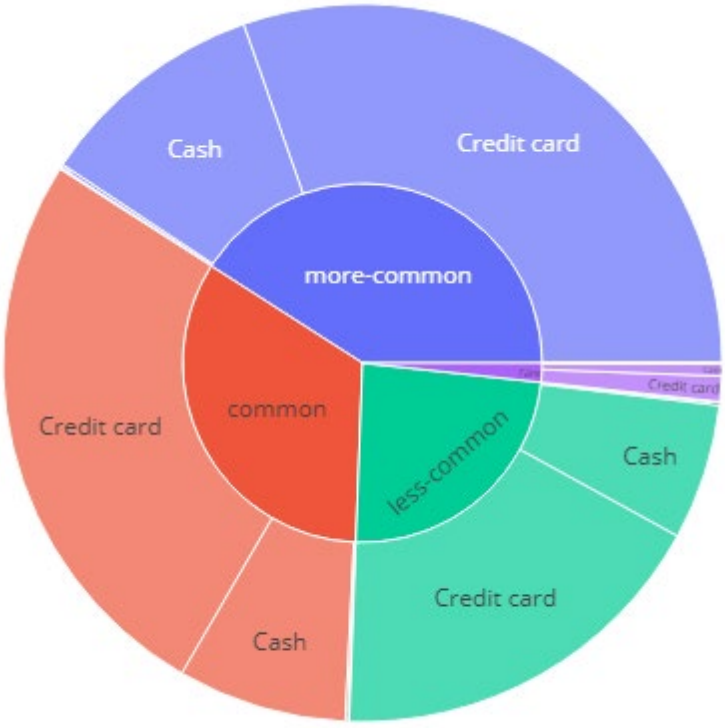
3.1.3 طرق الدفع المستخدمة

معظم طرق الدفع لسيارات الأجرة في نيويورك تتم عن طريق ال credit card.



وذلك يعود إلى أن كلتا الوكالتين هي من الوكالات التكنولوجية الحديثة كما أنه وبعد إجراء القليل من البحث تبيّن لنا أنّ أغلب سائقي سيارات الأجرة في مدينة نيويورك يعتمدون الـ credit cards كوسيلة دفع.

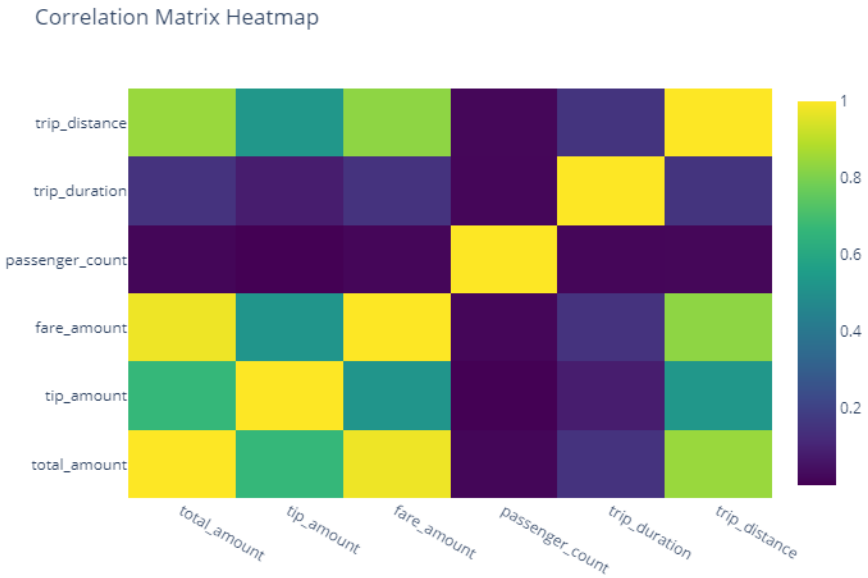
3.1.4 طرق الدفع لكل صنف من أصناف الرحلة



3.1.5 الارتباط الخطي بين الكلفة الإجمالية ومبلغ الإكرامية ومقدار الأجرة وعدد الركاب ومدة الرحلة والمسافة المقطوعة

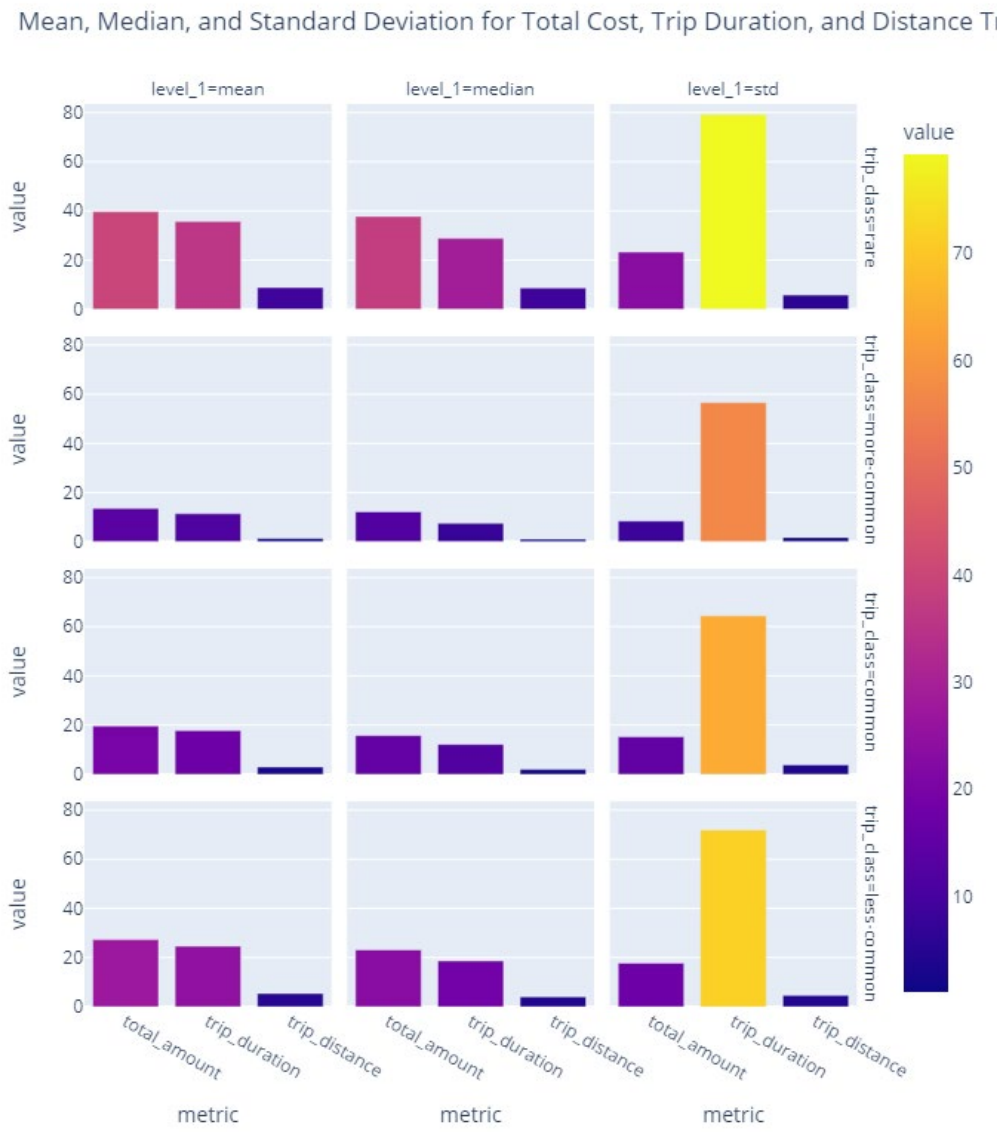
نلاحظ أن المسافة مرتبطة جداً بالأجرة الكلية والأجرة الفعلية، وبالطبع الأجرة الفعلية مرتبطة جدا بالأجرة الكلية. والعلاقة بين عدد الركاب وباقي السمات ضعيف (لأنه ليس لديهم تكسي ركاب 😬).

العلاقة بين المسافة ومدة الرحلة ليست مرتبطة جداً لأن هناك عوامل خارجية قد تؤثر على مدة الرحلة (السرعة، الازدحام، المسافات الطويلة غالباً تكون رحلات سفر وحدود السرعة فيها عالية لخلوها من الازدحام الشديد، ...).



3.1.6 المقارنة بين المتوسط والوسيط والانحراف المعياري للكلفة الإجمالية ومدة الرحلة والمسافة المقطوعة من أجل كل صنف من أصناف الرحلة

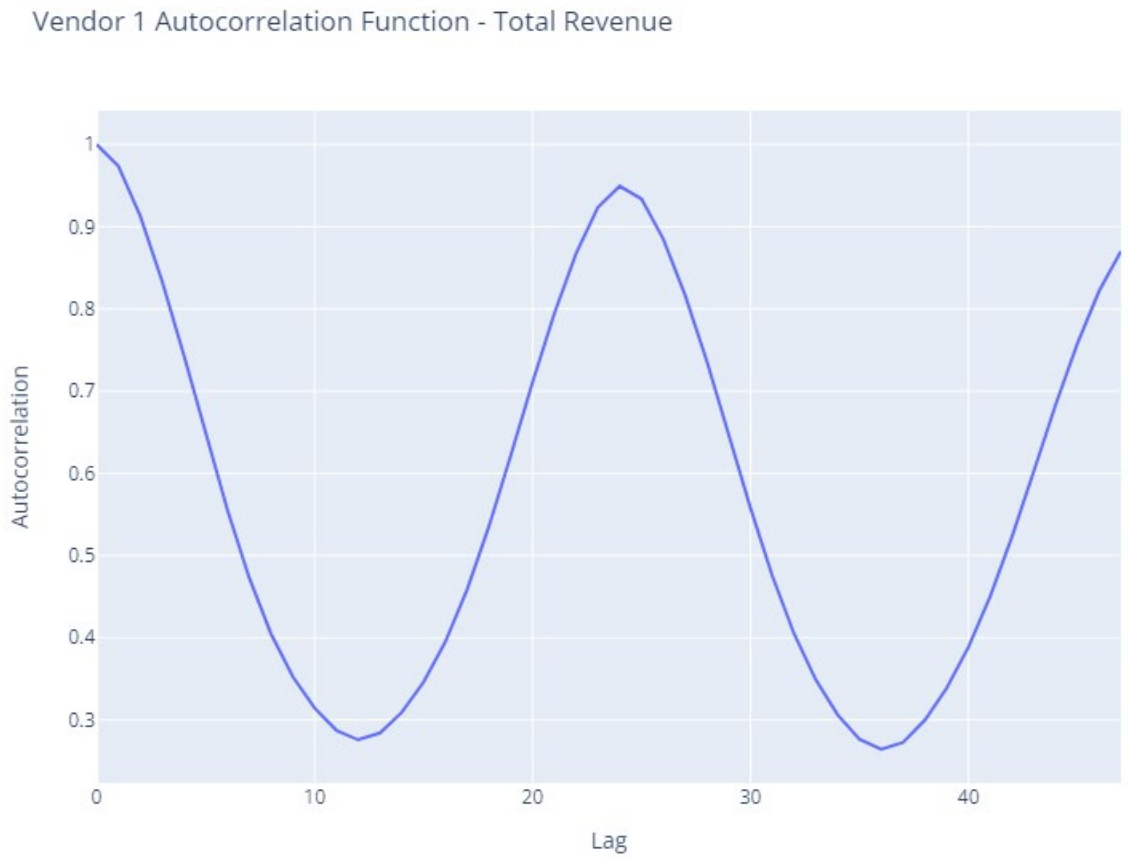
نلاحظ أن مدة الرحلة انحرافها عالي جداً لنفس السبب المذكور سابقاً (العوامل الخارجية المؤثرة بمدة الرحلة)

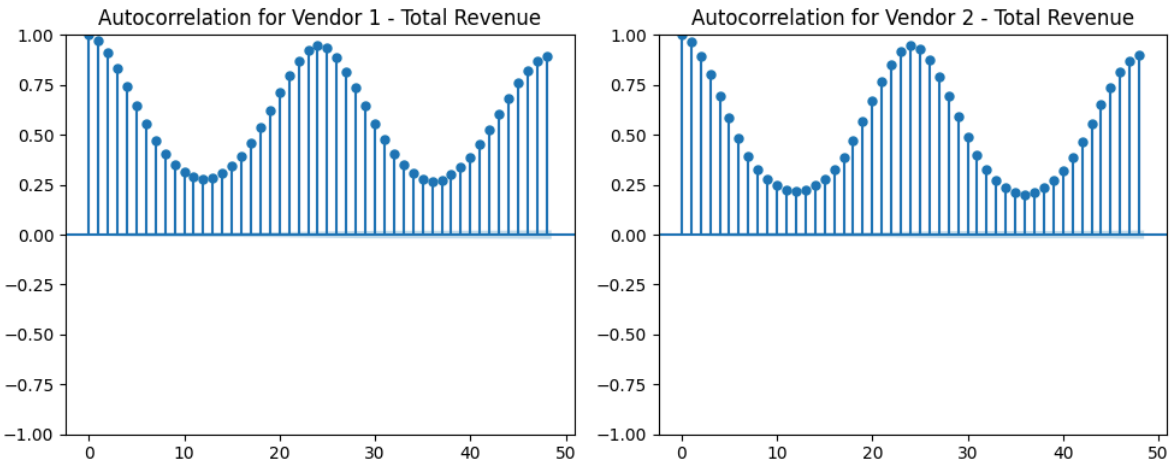


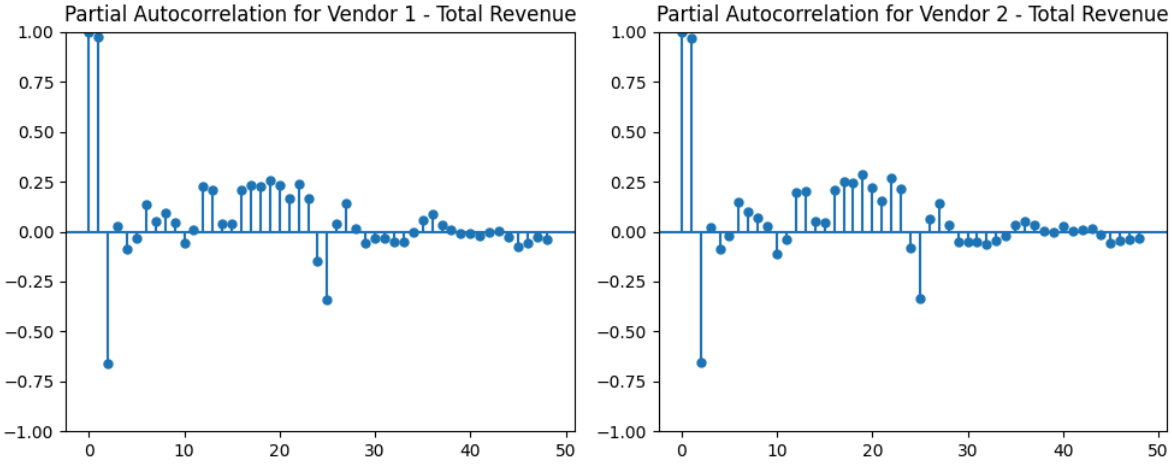
من المخطط السابق نلاحظ ما يلي:

1- الانحراف المعياري لمدة الرحلة بالنسبة للرحلات النادرة عالي جداً.
2- متوسط ووسيط وانحراف مسافات الرحلات الأكثر تكراراً منخفض جداً وهذا يعني أنه تقريباً جميع المسافات ذات قيم صغيرة.
3- أكبر متوسط لـ total_amount كان من أجل الرحلات النادرة وقيمته عالية نسبياً.
4- إن متوسط الـ total_amount بأصغر قيمة له من أجل الرحلات الـ more common.
5- يمكن اعتبار أنّ الانحراف المعيار للمسافة قليل من أجل كل أنواع الرحلات 🧐

3.1.7 ترابط السلاسل الزمنية Partial autocorrelationg Autocorrelation







عند رسم مخطط ACF على 48 lags لاحظنا أنه هناك seasonality في time series من أجل الإيرادات للوكالة الأولى وهذا يعني أن الإيرادات تتكرر تقريباً كل 24 ساعة أي تكون الإيرادات متقاربة في الساعات نفسها من كل يوم.

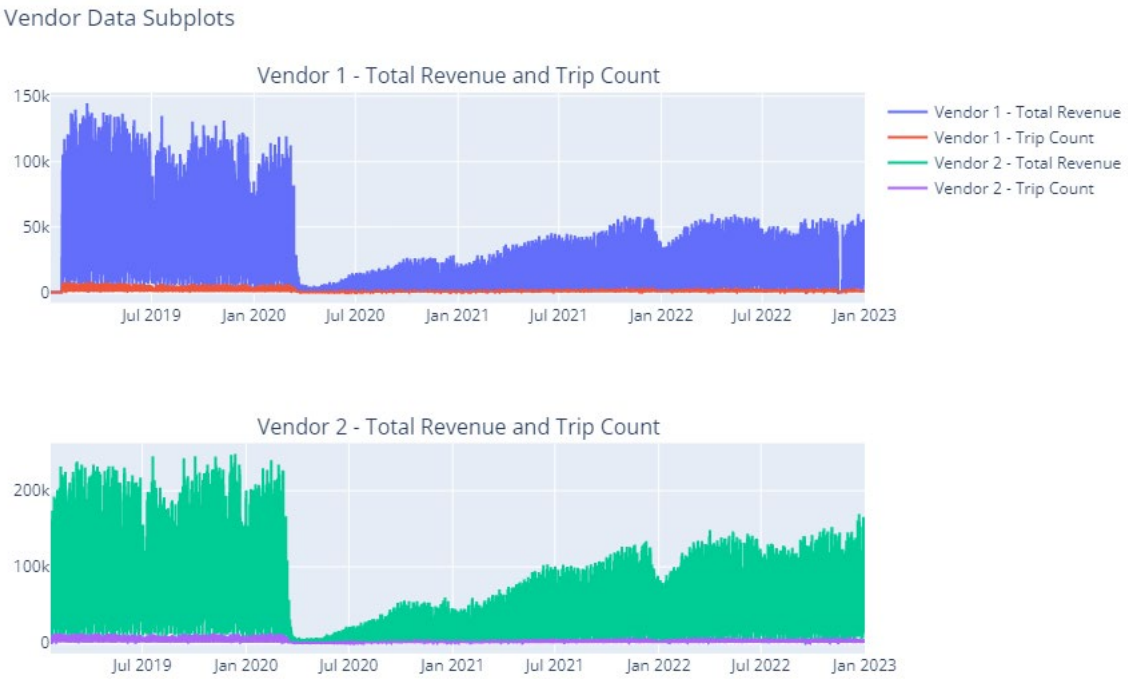
وعند رسم مخطط PACF لاحظنا قيم كبيرة بالترابط الجزئي من أجل lag=1 و lag=2.

3.1.8 نسخة ملساء من كل سلسلة باستخدام Rolling Window

نافذة من أجل السنين: وهنا يكون الانخفاض ملحوظاً بالشكل الأكبر.

نافذة من أجل الأشهر:

3.1.9 عمليات تحليل إضافية



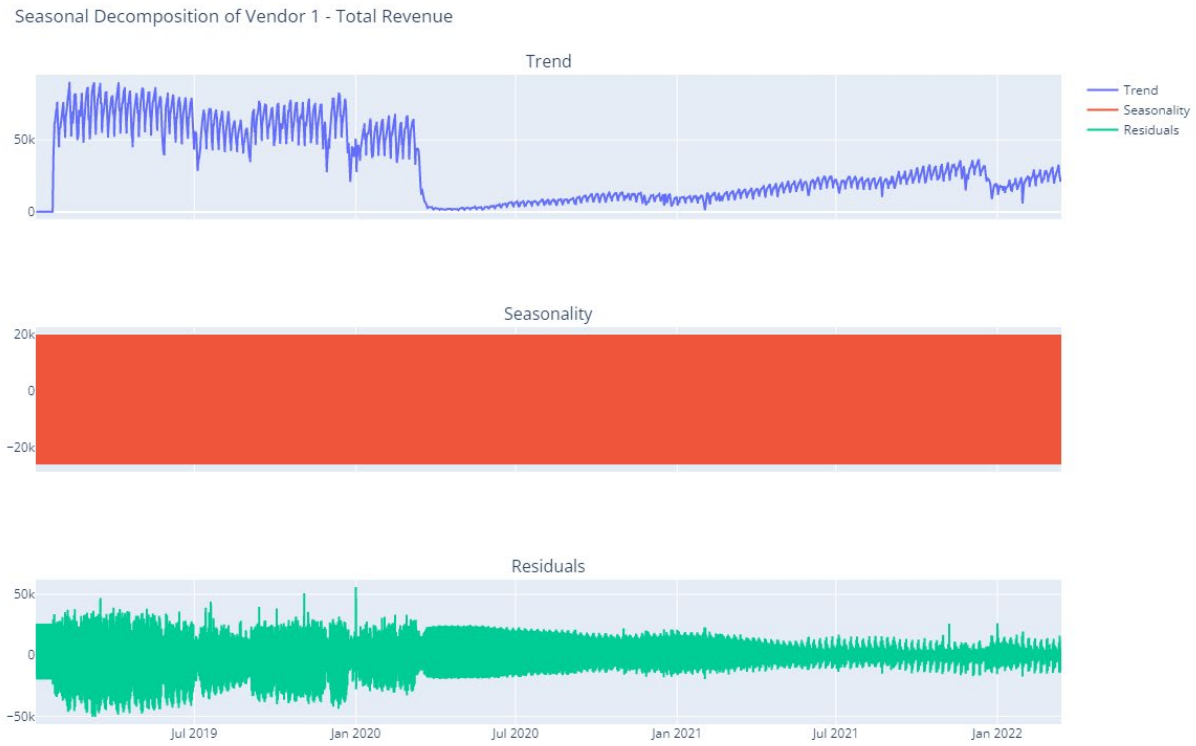
قمنا هنا برسم السلسلة الزمنية على مدار شهري، ولاحظنا أنّ هناك انخفاض ملحوظ بعدد الرحلات والإيرادات يتبعه ارتفاع بنسبة قليلة من أجل كلتا الوكالتين ما بين شهري نيسان 4 من عام 2020 وتشرين الأول 9 من عام 2020، وبعد إجراء بعض الأبحاث وجدنا أن فترة فيروس كورونا COVID-19 قد أثرت تأثيراً سلبيّاً كبيراً على وكالات وسائقي سيارات الأجرة بشكل عام حتى أنهم في ايرلاند خرجوا باحتجاجات، وبحسب الإحصائيات فإن أكثر من نصف سائقي سيارات الأجرة لم يعودوا إلى عملهم بعد الجائحة وهذا يفسر أنه حتى بعدما عادت الإيرادات وعدد الرحلات لترتفع من جديد إلا أنها لم تعد كما كانت عليه قبل فترة الجائحة.

3.2 اصطياد الأنماط

3.2.1 نمذجة مكونات السلاسة الزمنية لكل وكالة باستعمال نموذج إحصائي مناسب، وبناء النموذج النهائي باستخدام prophet، مع عمليات التوليف والمقارنة بين النماذج

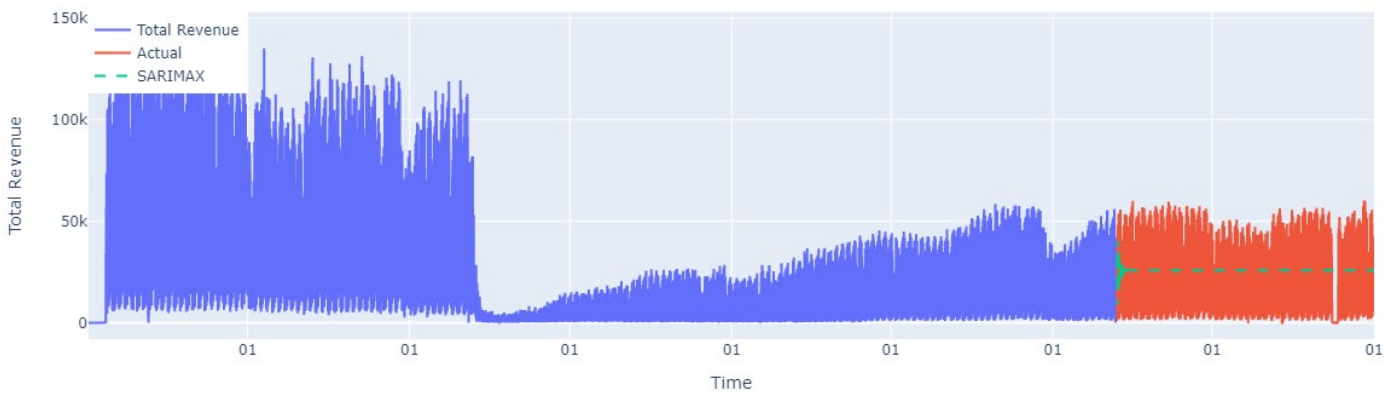
بدايةً، يمكن بشكل عام ملاحظة أن الـ time series تحوي seasonality وللتأكد سنقوم بعمل decomposition وهذه النتائج:

vendor 1:



عند تكبير المخطط بالنسبة لـ seasonality نلاحظ بوضوح ان الداتا seasonal.

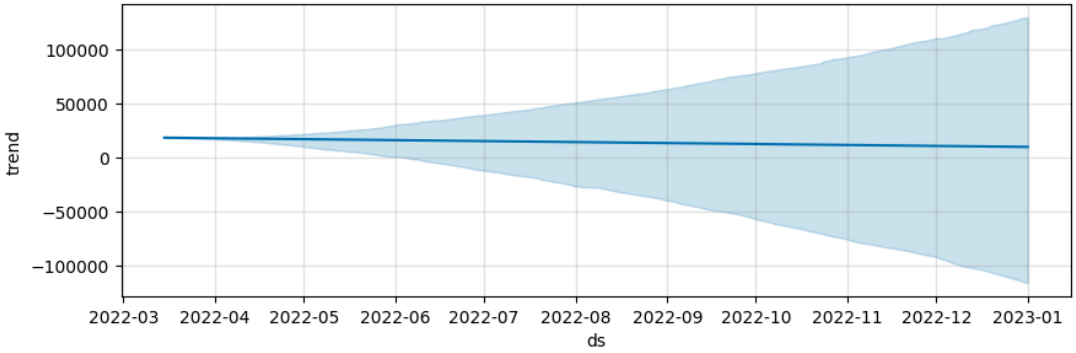
لذلك قمنا باختيار SARIMAX كمودل أول لتجريبه. وبعد العديد من عمليات fine-tuning اخترنا القيم التالية: p: 6 d: 1 q: 3.

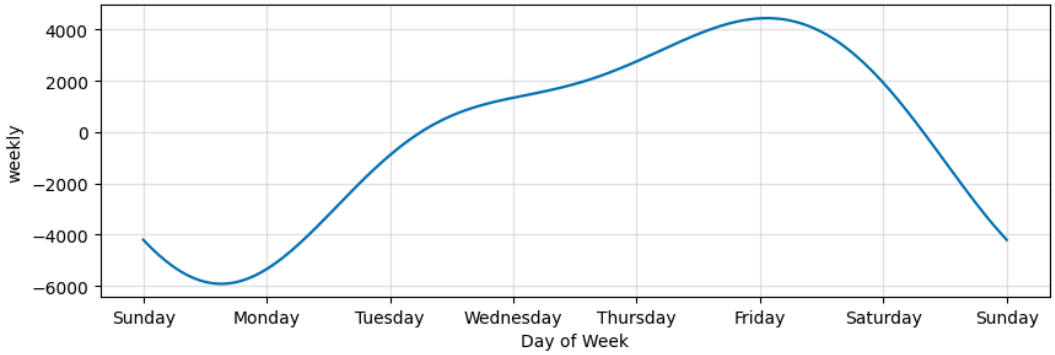


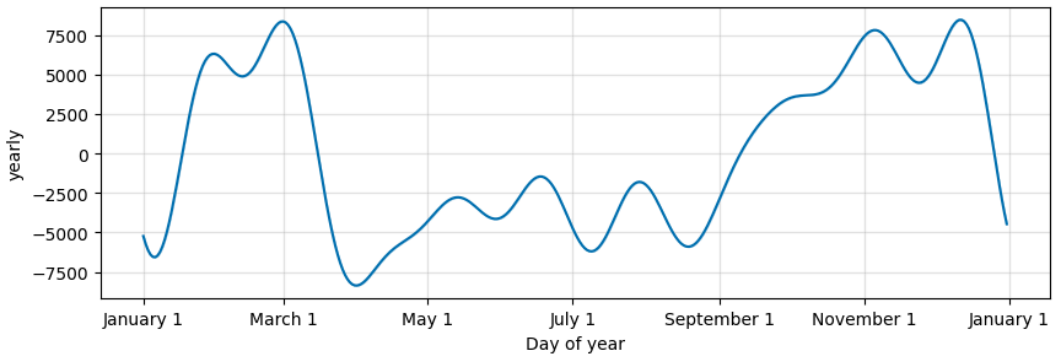
بناء النموذج النهائي باستخدام prophet:

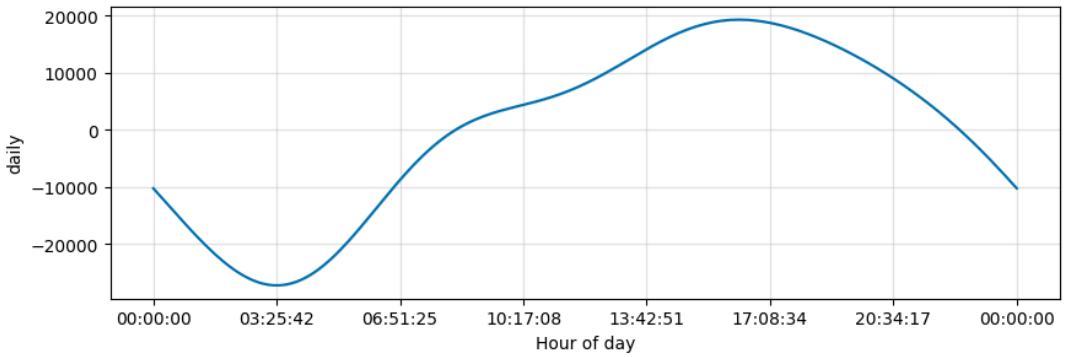
Vendor1:

Prophet Components:

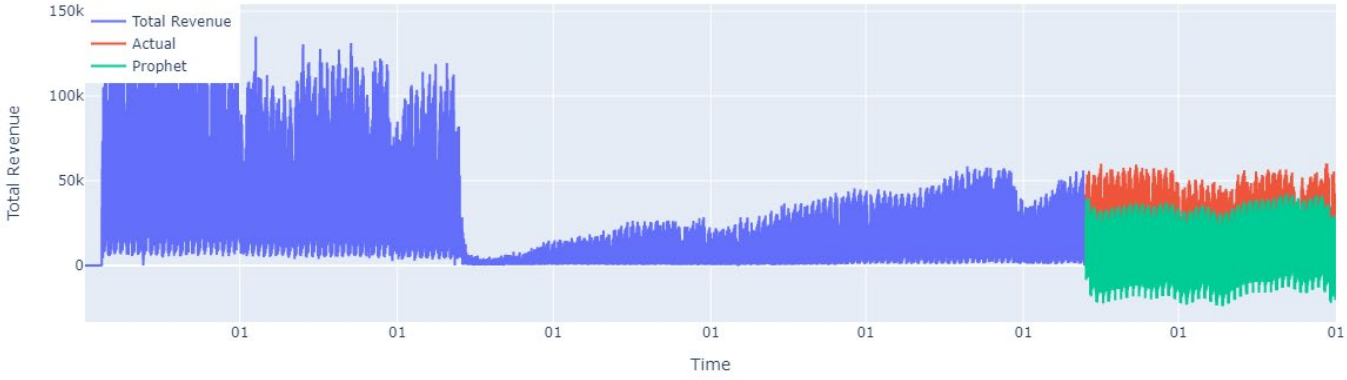






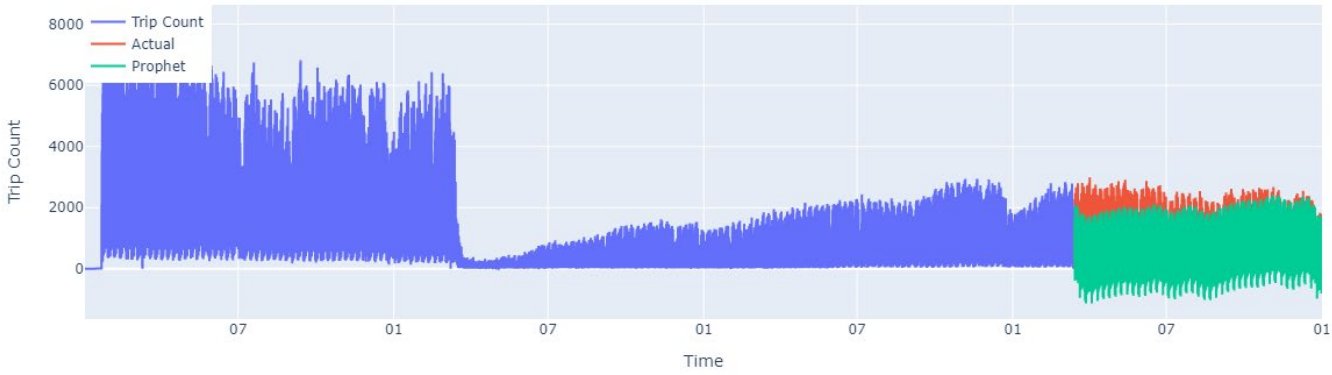


توقّع الـ prophet:

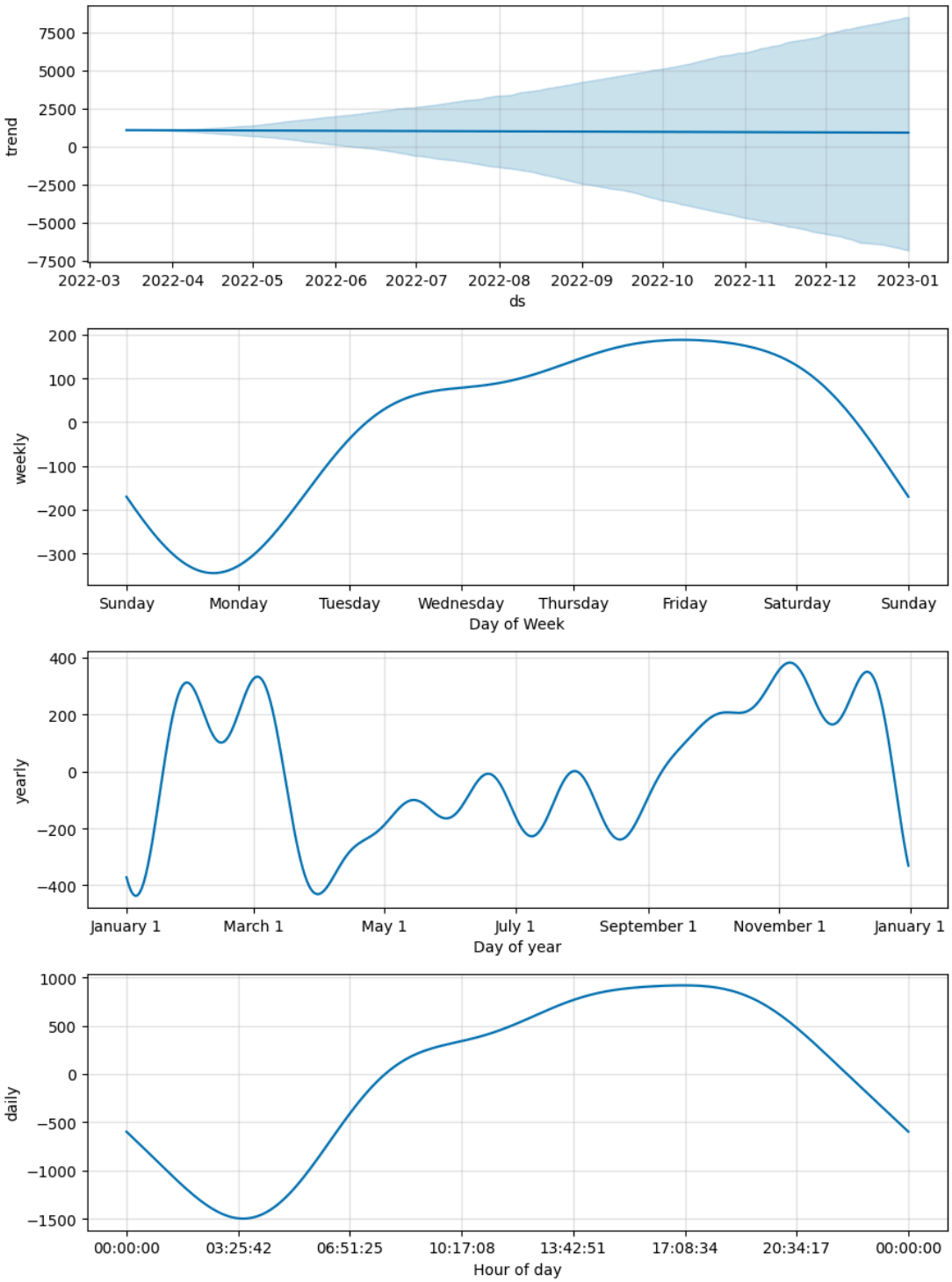


Vendor2:

توقّع الـ prophet:



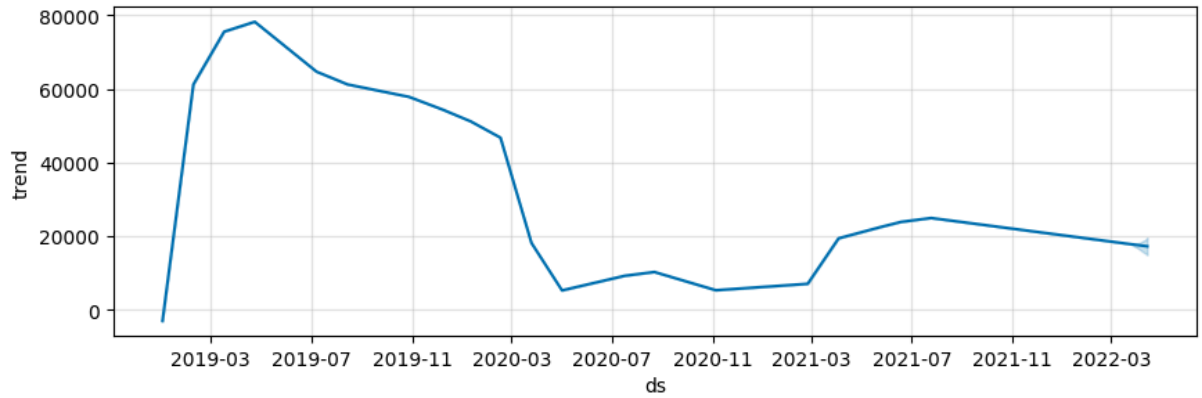
Prophet Components:

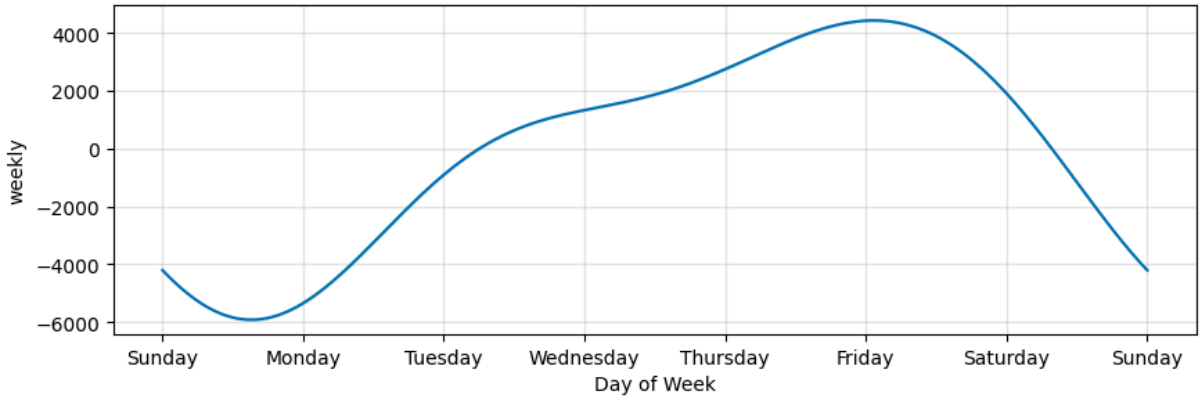


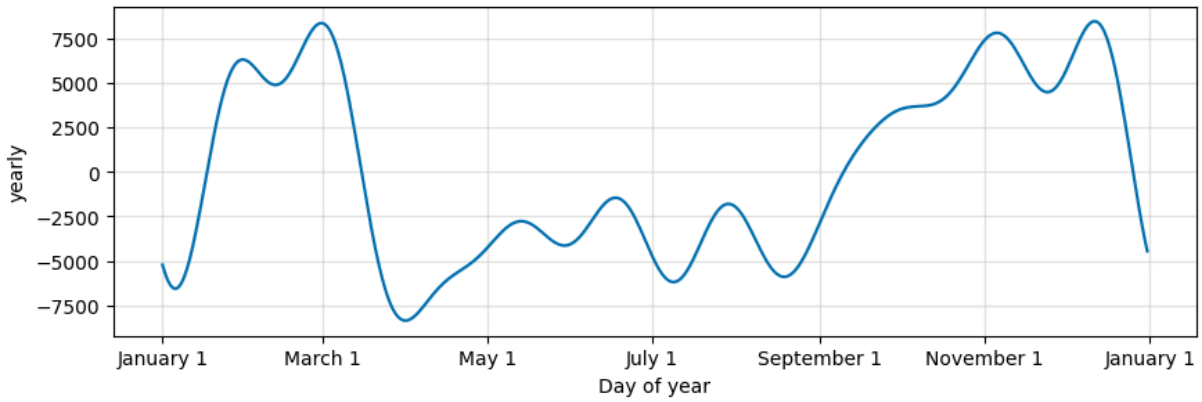
3.3 الوقت قيم

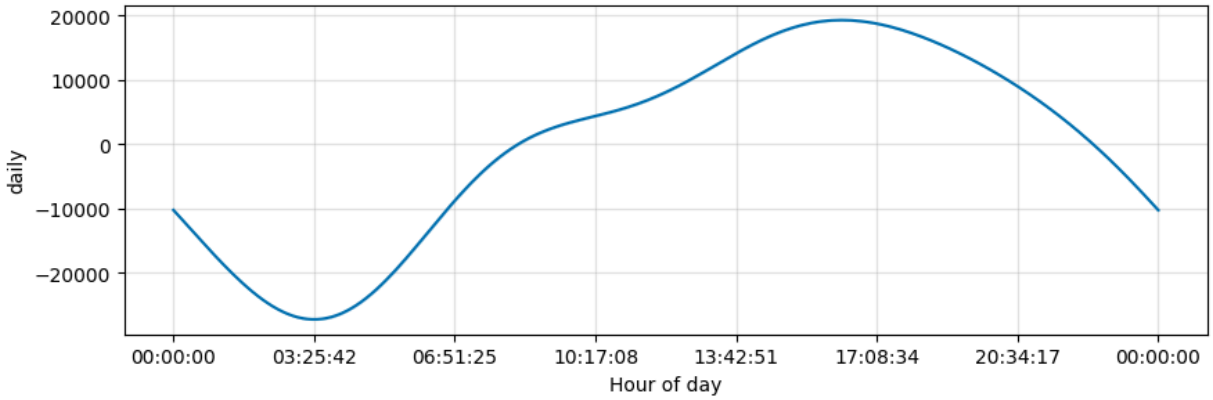
3.3.1 التنبؤ بالطلب المستقبلي لرحلات وإيرادات سيارات الأجرة، مع المقارنة والتقييم باستخدام المقاييس المناسبة، وتفسير النتائج في سياق البيانات.

نلاحظ أنّ من أجل trend فإن التنبؤ بالطلب المستقبلي يتبع الأنماط بشكل جيد









ويكررها من أجل التنبؤ.

3.4 عشوائية القصص

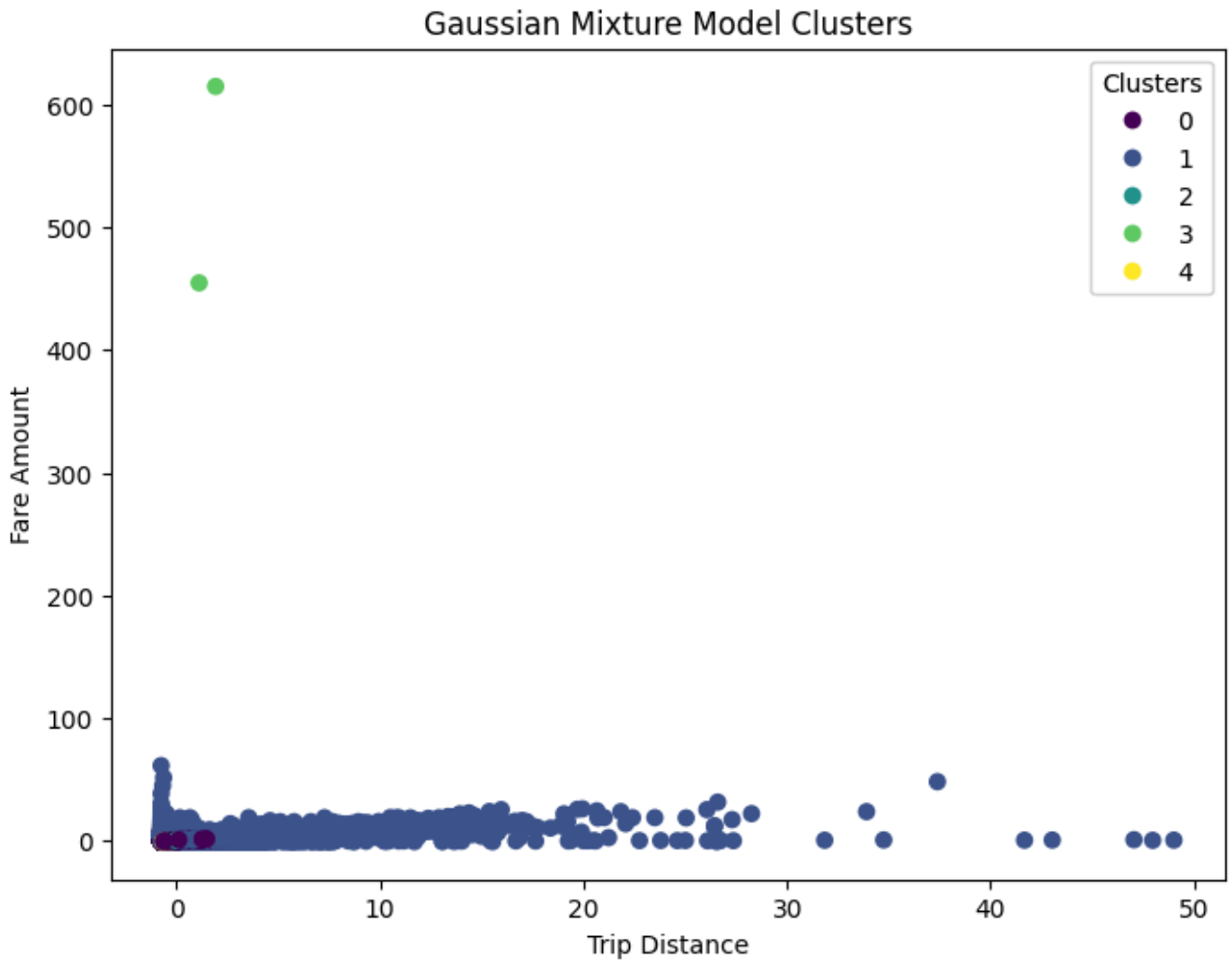
3.4.1 Sampling للبيانات

جرّبنا إجراء عملية sampling عشوائية وأخذنا قيمة 10% وهي القيمة الإحصائيّة المعتمدة عادةً.

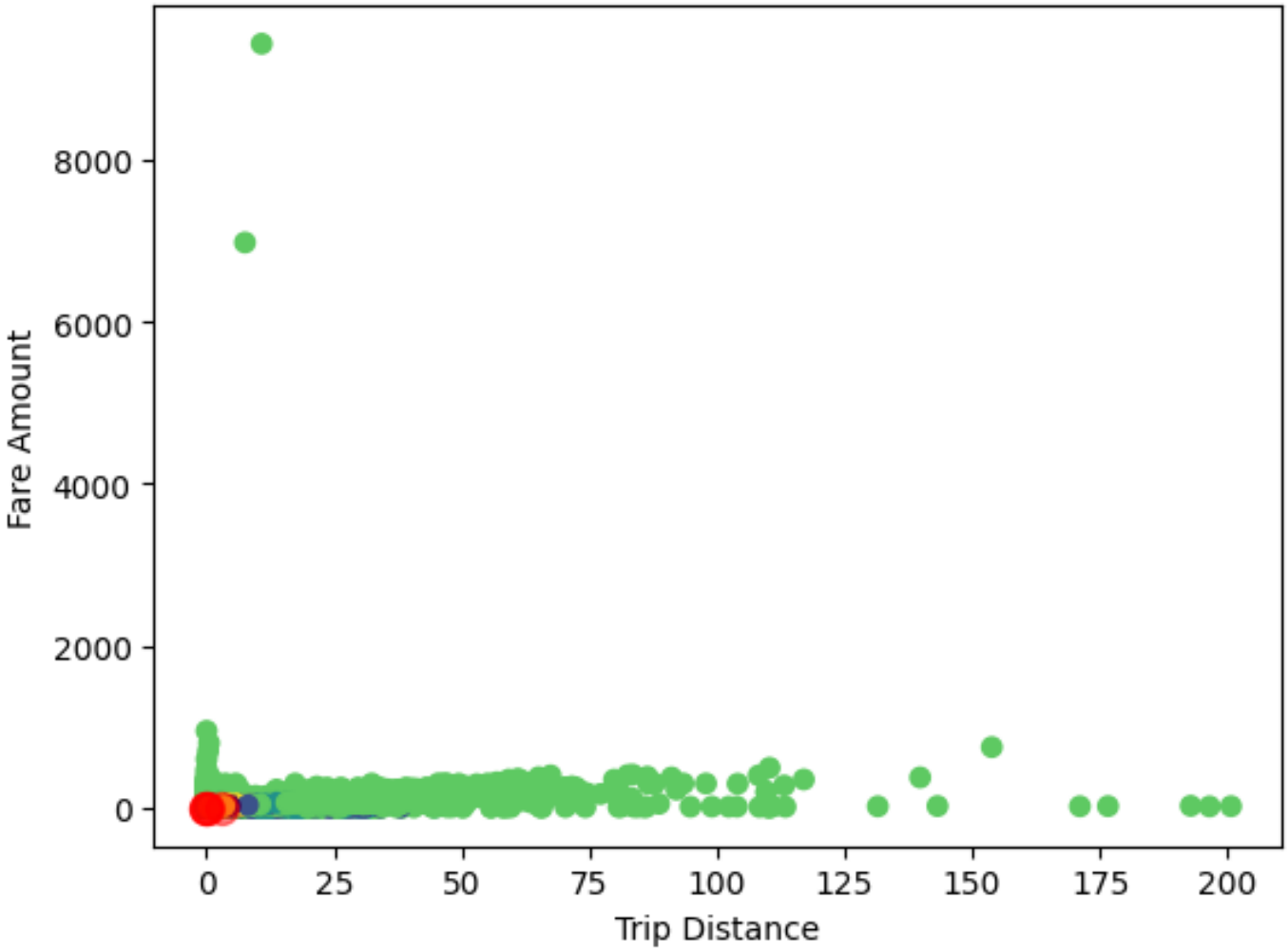
كذلك جربنا stratified sampling وذلك بعد تقسيم الداتا لمجموعات تتبع لسمة جديدة هي ما قبل كورونا وخلال كورونا وما بعد كورونا وأخذ عينا عشوائية من أجل كل مجموعة بنسبة 10% ولكن لم نستطع أن نتابع طلب clustering بالنسبة لهذا الـ sampling.

3.4.2 Clustering

Gaussian mixture model:



K-means clustering:



4 المراجع:

How Curb Became a Multimillion-Dollar App by Bringing Taxis Into the Uber Economy (uschamber.com)

About Us | Curb (gocurb.com)

CMT Group | Solutions for the Industry by the Industry

Manhattan - New York's most important and popular borough (introducingnewyork.com)

Does NY City Cabs take credit card or just cash ? - New York City Forum - Tripadvisor

Taxi driver shortage prompts public safety fears - BBC News

'Killing an industry that's already dead': Taxi drivers stage protest in Dublin city (thejournal.ie)