كلية الهندسة المعلوماتية

جامعة دمشق

يوميات تأجير الدراجات

وظيفة تحليل البيانات

إعداد

أحمد بسام عبود

معاذ الصبح

كنانة الغزالي

مقدمة: تمهيد الطريق لاستكشاف البيانات

تبدأ رحلتنا في عالم بيانات تأجير الدراجات بالإعدادات الأساسية, لتحضير بيئة التحليل وتحميل البيانات الخام. تضمن هذه الخطوة التأسيسية توفر جميع الأدوات اللازمة وجاهزية مجموعة البيانات للفحص الدقيق.

استيراد الأدوات: المكتبات الأساسية

قبل الغوص في البيانات, قمنا بتجهيز مساحة عملنا بمجموعة من مكتبات بايثون القوية. هذه المكتبات هي أساس جهودنا في تنقيب البيانات, أهمها:

- "scipy.spatial.cKDTree": تستخدم للاستعلامات المكانية الفعالة, خاصة في العثور على أقرب الجيران, وهو أمر بالغ الأهمية للتحليل المستند إلى الموقع
- geopy.distance.geodesic: تستخدم لحسابات المسافة الجيوديسية الدقيقة بين الإحداثيات الجغرافية, مما يعزز فهمنا لأطوال الرحلات والعلاقات المكانية.
- Geopandas: مكتبة متخصصة للعمل مع البيانات الجغرافية المكانية, تمكننا من إجراء العمليات على الميزات الجغرافية.

تحميل مجموعة البيانات الأساسية: رحلات تأجير الدراجات اليومية

يُشكل ملف "daily-rent-sampled.parquet" جوهر تحليلنا, حيث يحتوي على سجلات مفصلة لرحلات تأجير الدراجات اليوميةكما انه لاحقا سيتم دمج اغلب الاعمدة في هذا الملف. توفر مجموعة البيانات هذه عرضاً دقيقاً لكل رحلة, مع أعمدة تصف جوانب مختلفة من الرحلة. تُظهر النظرة السريعة للصفوف القليلة الأولى بنية البيانات ومحتواها.

فيما يلي تفصيل للأعمدة الهامة ضمن مجموعة البيانات هذه:

"ride_id" مُعرف فريد لكل رحلة دراجة.

"rideable_type" نوع الدراجة المستخدمة (على سبيل المثال, دراجة كلاسيكية, دراجة كهربائية).

"started_at" تاريخ بدء الرحلة.

"ended_at" تاريخ انتهاء الرحلة.

"start_station_name" اسم المحطة التي بدأت منها الرحلة.

"start_station_id" مُعرف فريد للمحطة البدائية.

"end_station_name" اسم المحطة التي انتهت عندها الرحلة.

"end_station_id" مُعرف فريد للمحطة النهائية.

"start_lat", "start_lng" إحداثيات خط العرض والطول للمحطة البدائية, على التوالي.

"end_lat", "end_lng" إحداثيات خط العرض والطول للمحطة النهائية, على التوالي.

"member_casual" تصنيف المستخدم كعضو مسجل أو راكب عابر.

فحص البيانات:

قبل أن نتمكن من استخلاص رؤى ذات معنى من بياناتنا, يجب علينا أولاً فهم خصائصها الكامنة. يتضمن ذلك فحصًا شاملاً لبنيتها ومحتواها ووجود أي قيم مفقودة.

- **نظرة عامة أولية على الهيكل:**

  تتكون مجموعة بياناتنا, "df_rent", من 2,018,456 (صف) و 13 عمودًا. يوفر هذا الحجم الكبير أساسا غنيا للتحليل. تتكون الأعمدة من مزيج من أنواع البيانات:

  - تاريخ ووقت عمودان "started_at" :g"ended_at" , وهما حاسمان لتحليل السلاسل الزمنية.
  - عشري (4 أعمدة)"end_lng", "end_lat", "start_lng", "start_lat" :, تمثل الإحداثيات الجغرافية.
  - كائن (7 أعمدة): " ride_id", "rideable_type", "start_station_name", "start_station_id", "member_casual", "end_station_name", "end_station_id", تمثل بيانات فئوية أو معرفة.

- **لمحة إحصائية عن البيانات الرقمية:**

  يكشف الملخص الإحصائي لأعمدتنا الرقمية (" started_at", "ended_at", "start_lat", "start_lng", "end_lat", "end_lng") عن أنماط مثيرة للاهتمام. على سبيل المثال, تقع تواريخ الرحلات إلى حد كبير ضمن عام 2024, مع بعض القيم الشاذة في (started at) الحد الادنى لموعد البدأ 1970و الحد الاعلى لموعد الانتهاء (ended_at) في 2030 والتي سنركز عليها اثناء تنظيف البيانات لازالة الرحل غير المنطقية (مثل الرحل التي تبدا في المستقبل وتنتهي في الماضي والرحل التي مدتا سنة) هذه رحل غير منطقية ويجب ازالتها قبل البدء بدراسة معمقة للبيانات.

- **الكشف عن القيم المفقودة:**

تعد خطوة حاسمة في تقييم جودة البيانات هي تحديد القيم المفقودة. يكشف تحليلنا عن عدة أعمدة تحتوي على بيانات مفقودة:

- "start_station_name" 393,290 قيمة مفقودة(19.48%)
- "start_station_id" 393,290 قيمة مفقودة(19.48%)
- "end_station_name" 408,636قيمة مفقودة(20.24%)
- "end_station_id" 409,167 قيمة مفقودة(20.27%)
- "end_lat" 1,552قيمة مفقودة(0.08%)
- "end_lng" 1,552 قيمة مفقودة (0.08%)

- **تصنيف الأعمدة:**

لنهج منظم لمعالجة البيانات, قمنا بفصل أعمدتنا إلى أنواع رقمية وفئوية:

- الأعمدة الرقمية"start_lat", "start_lng", "end_lat", "end_lng"
- الأعمدة الفئوية "ride_id", "rideable_type", "start_station_name", "start_station_id", "end_station_name", "end_station_id", "member_casual"

بالإضافة إلى ذلك, تم تحديد "rideable_type" و "member_casual" كأعمدة فئوية وصفية, مما يعني أنها قد تكون مفيدة بشكل خاص للتحليل المباشر والتصور دون تحويلات واسعة النطاق.

- **تنظيف البيانات والمعالجة المسبقة:**

مع فهمنا للحالة الأولية لبياناتنا, تتضمن المرحلة الحاسمة التالية التنظيف والمعالجة المسبقة. يضمن ذلك جودة البيانات واتساقها وملاءمتها للتحليل المتعمق والنمذجة يتضمن التنظيف تعبئة القيم المفقودة بطريقة مناسبة او ازالتها.

- **التعامل مع التناقضات:**

نتعامل مع التناقضات المحتملة في مدة الرحلة عن طريق إزالة الرحلات التي يكون فيها "ended_at" أسبق من "started_at", أو عندما تكون "duration_min" سالبة أو طويلة بشكل مفرط (على سبيل المثال, أكثر من 1440 دقيقة, أي 24 ساعة). يضمن هذا التصفية أن تكون حسابات المدة منطقية وذات صلة.

كما لاحظنا بعض الطوابع الزمنية المتطرفة لـ "started_at" و "ended_at" (على سبيل المثال, تواريخ في 1970 أو 2000, وتواريخ بعيدة في المستقبل مثل 2030). للتركيز على البيانات ذات الصلة والحديثة, قمنا بتصفية مجموعة البيانات لتشمل فقط الرحلات التي بدأت وانتهت خلال عام 2024. يضمن هذا التحسين الزمني أن تحليلنا يهتم برحلات هذه السنة فقط.

دمج البيانات الخارجية:

للحصول على فهم أكثر شمولاً, قمنا بدمج مجموعات بيانات خارجية:

- ❖ بيانات الطقس: معلومات الطقس بالساعة لواشنطن العاصمة من 1 يناير 2024 إلى 31 ديسمبر 2024, بما في ذلك "conditions" و "humidity" و "windspeed" و "temp". يسمح لنا ذلك بالتحقيق في تأثير الطقس على استخدام الدراجات.
- ❖ بيانات المحطات: معلومات حول محطات Capital Bikeshare مع ملف "stations.csv" بما في ذلك سعات المحطات
- ❖ مناطق وقوف السيارات: باستخدام "Residential_and_Visitor_Parking_Zones.geojson", قمنا بتحديد ما إذا كانت نقطة بدء أو نهاية الرحلة تقع ضمن منطقة سكنية, مما أدى إلى إنشاء "start_parking_zone_name" و "end_parking_zone_name".
- ❖ منطقة الأعمال المركزية (CBD): استخدمنا "DDOT_Central_Business_District.geojson" لتحديد ما إذا كانت نقطة بدء أو نهاية الرحلة تقع ضمن منطقة الأعمال المركزية, مما أدى إلى إنشاء علامات منطقية "start_in_cbd" و"end_in_cbd".

- القرب من وسائل النقل العام: باستخدام "Metro_Bus_Stops.csv" و "Shuttle_Bus_Stops.csv", قمنا بحساب المسافة الجيوديسية إلى أقرب محطات للحافلات والمترو لكل من مواقع البدء والانتهاء, مما أدى إلى إنشاء ميزات مثل "start_nearest_bus_stop_distance_m", "start_nearest_metro_stop_distance_m", "end_nearest_bus_stop_distance_m", "end_nearest_metro_stop_distance_m".g
- المسافة إلى منطقة الأعمال المركزية وعلامة القرب: قمنا بحساب "end_distance_to_cbd_m"ثم أنشأنا ميزة فئوية "end_proximity_to_cbd" للإشارة إلى ما إذا كان موقع النهاية 'قريب' أو 'بعيد' عن منطقة الأعمال المركزية.

حساب التكلفة:

ميزة بسيطة هي "trip_total_cost", محسوبة بناءً على "duration_min" و "member_casual". يسمح هذا بتحليل واقعي للإيرادات.

تصنيف سعة المحطة

لفهم أفضل لتأثير حجم المحطة, قمنا بتصنيف سعات المحطات إلى 'صغيرة' و 'متوسطة' و 'كبيرة' بناءً على الفئات المحددة مسبقًا. وقد أدى ذلك إلى إنشاء "start_station_capacity_category" و "end_station_capacity_category".

التجزئة الجغرافية ونشاط المنطقة

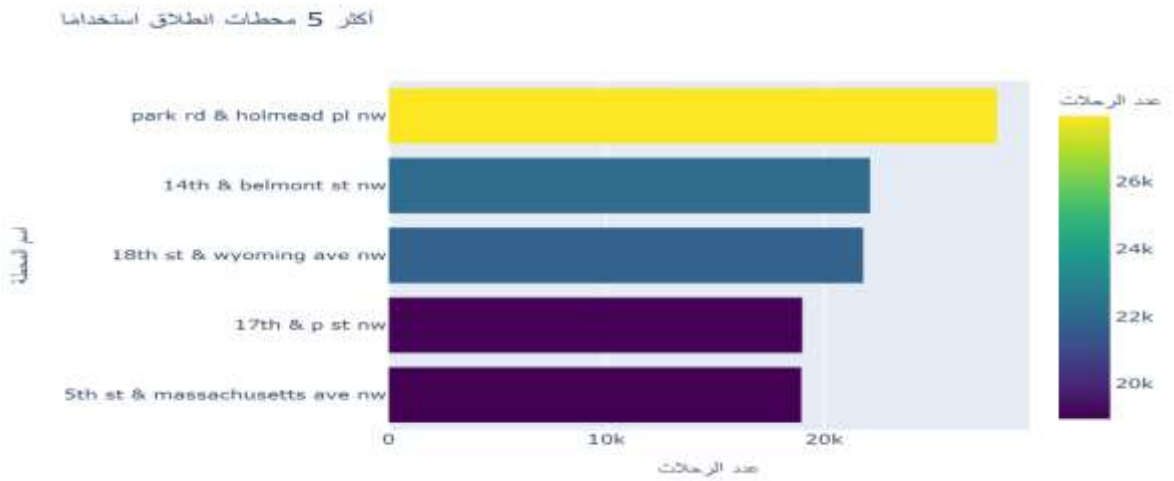
لتحليل مكاني أكثر تقدما, قمنا بإنشاء تجزئات جغرافية لمواقع البدء والانتهاء (" ,"start_geohash "end_geohash"). وقمنا أيضا بحساب "start_zone_activity", الذي يمثل العدد الإجمالي لرحلات الدراجات التي تنشأ من كل منطقة بدء.

إعداد البيانات النهائية للنمذجة

لأغراض التنبؤ, قمنا بتجميع بيانات الإيرادات اليومية, بما في ذلك إجمالي التكلفة, هذا يهيئ البيانات لنمذجة السلاسل الزمنية باستخدام Prophet

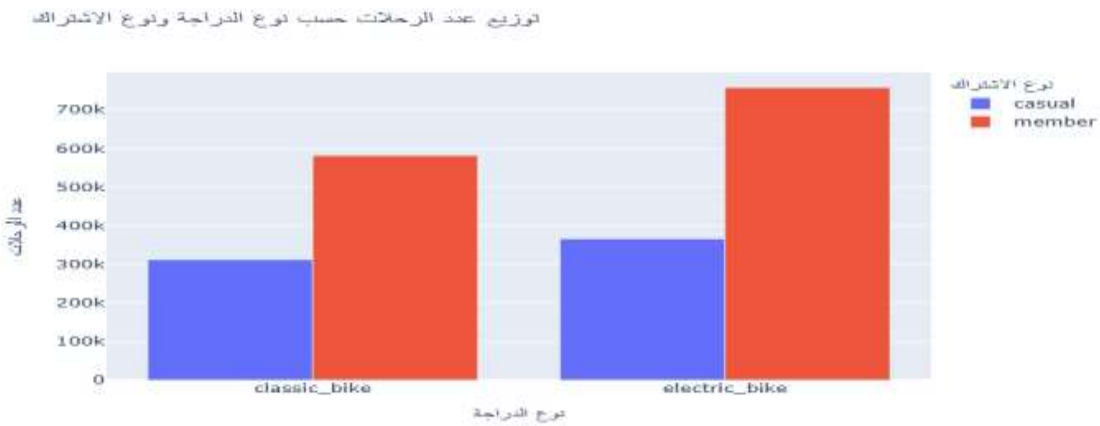
تحليل البيانات الاستكشافي

5.1.أكثر خمس محطات انطلاق استخدامًا



تم استنتاج الشكل عن طريق حساب عدد الرحلات المنطلقة من كل محطة ونستنتج من المخطط ان محطة "بارك رود وهولميد بي إل شمال غرب" نقطة جذب رئيسية لرحلات الدراجات, وقد تكون موقعًا استراتيجيًا لصيانة الدراجات أو إعادة توزيعها أو حملات التسويق لانها اكثر مراكز الانطلاق استخداما .

5.2.توزيع الرحلات حسب نوع الدراجة ونوع الاشتراك

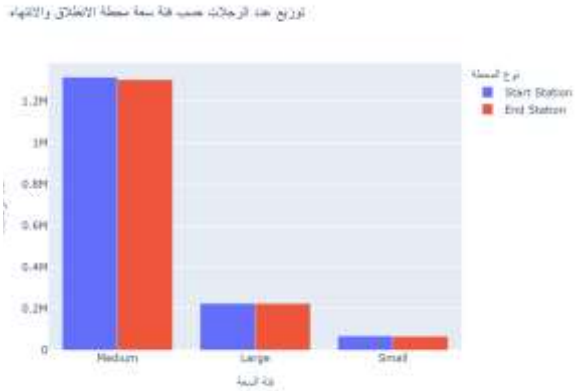


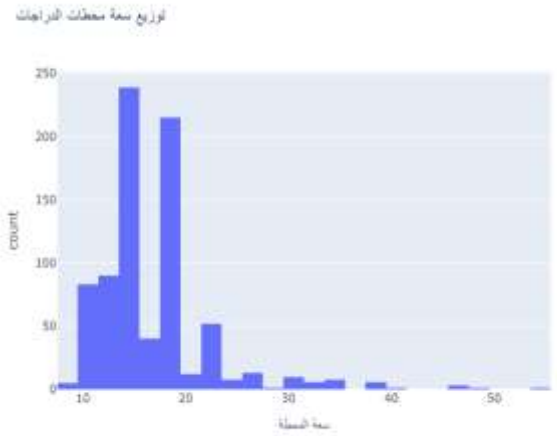
-نستنتج من المخطط ان الدراجات الكهربائية الخيار المُفضّل, خاصةً بين الأعضاء المُشتركين, وهذا استنتاج منطقي لان الدرجات الكهربائية عادة تتصف بالسرعة وسهولة القيادة كما انه نسب استعمال الدرجات الكلاسيكية ليس اقل بكثير بل يحظى بشعبية كبرى خاصة بين الاعضاء غير المشتركين

5.3.توزيع الرحلات حسب نوع الدراجة والاشتراك لأفضل 5 محطات بداية

نستنتج من هذا المخطط والذي يمكن اعتباره دمج ما بين الطلبين السابقين أن مستخدمي الدراجات الكهربائية الاعضاء هم المساهمون الرئيسيون في الرحلات من محطات الانطلاق الأكثر شيوعا. كما يستخدم الدراجون غير المنتظمين هذه المحطات أيضا, لكن يبدو أن إجمالي عدد رحلاتهم أقل من إجمالي عدد رحلات الأعضاء.

5.4.سعة محطات الدراجات

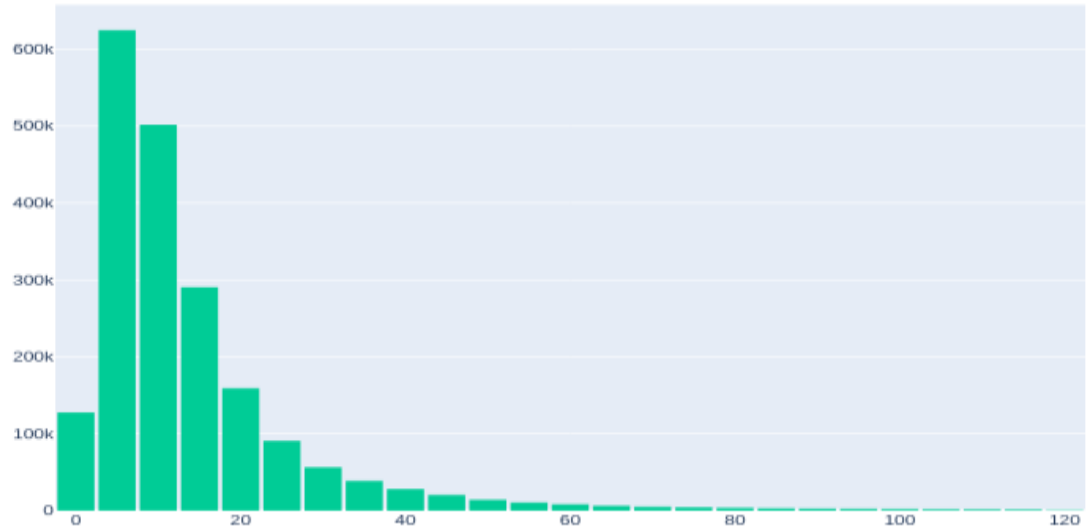




الشكل الاول: يشير هذا إلى أن نظام مشاركة الدراجات مُجهّز بشكل أساسي بعدد كبير من المحطات الصغيرة والمتوسطة الحجم,ونلاحظ ان العدد الاكبر للمحطات المتوسطة وهو توزيع منطقي بدلا من تقليل عدد المحطات وجعل سعتها كبيرة جدا التوزع المكاني لمحطات متوسطة وصغيرة يمكن ان يجني ارباحا اكثر .

الشكل الثاني: يشير هذا بقوة إلى أن المحطات "متوسطة الحجم" هي العمود الفقري لنظام مشاركة الدراجات, حيث تُدير الغالبية العظمى من الرحلات. يتوافق هذا مع ما توصل إليه الرسم البياني من وجود العديد من المحطات ذات السعة المتوسطة. تشير الأعداد المتساوية تقريبًا لنقاط الانطلاق والنهايات ضمن كل فئة إلى تدفق متوازن للدراجات.

5.5.رسم بياني لتوزيع مدة الرحلة

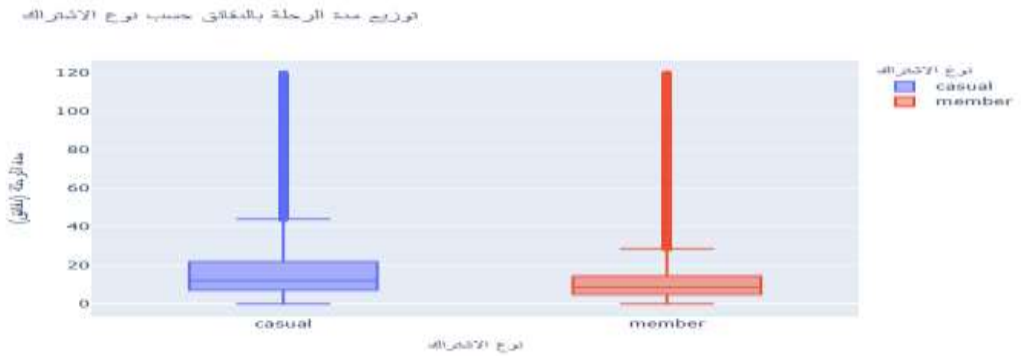


نستنتج من المخطط تركيز قوي للرحلات القصيرة, مع انخفاض حاد مع زيادة المدة .يشير هذا إلى أن غالبية تأجيرات الدراجات هي لرحلات سريعة وفعالة, ومن المحتمل أن تكون للتنقل أو السفر لمسافات قصيرة داخل المدينة .تسلط هذه الرؤية الأولية الضوء على حالة الاستخدام الرئيسية لخدمة مشاركة الدراجات

5.6.مدة الرحلة حسب نوع الدراجة

نستنتج من المخطط كيفية ارتباط تكلفة الرحلة بالمدة, بالتفريق بين الركاب العاديين والأعضاء, والدراجات الكلاسيكية مقابل الكهربائية .من الواضح أن المدد الأطول تؤدي بشكل طبيعي إلى تكاليف أعلى, ولكن توزيع النقاط عبر أنواع الركاب وأنواع الدراجات يقدم رؤى أعمق .على سبيل المثال, قد يقوم الركاب العاديون برحلات أطول وأكثر راحة, أو ربما يتكبدون تكاليف أعلى بسبب هياكل التسعير المختلفة .وبالمثل, قد يكون للدراجات الكهربائية, على الرغم من احتمال استخدامها لمسافات أطول, تسعير أو أنماط استخدام مختلفة مقارنة بالدراجات الكلاسيكية, مما يؤثر على العلاقة الإجمالية بين التكلفة والمدة.

5.7.رسم بياني لمدة الرحلة حسب نوع الاشتراك



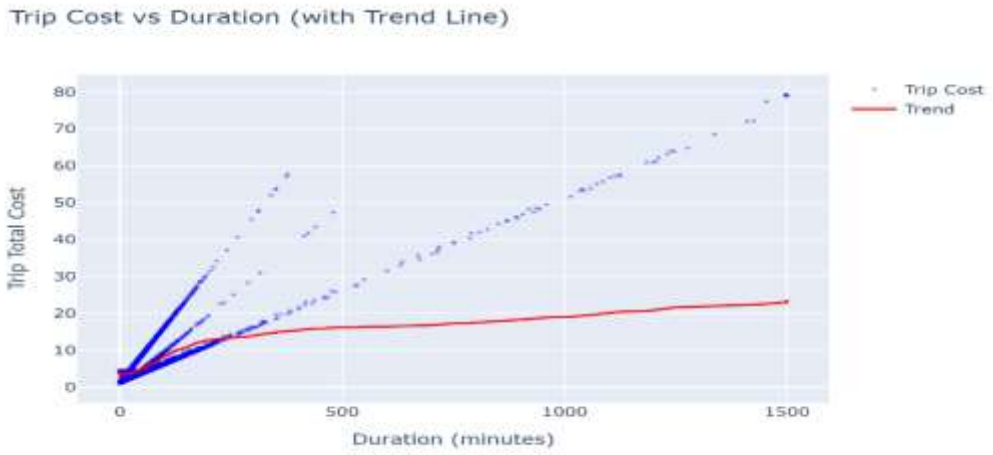
نستنتج من المخطط أن الأعضاء يستخدمون خدمة مشاركة الدراجات لرحلات قصيرة عملية وفعّالة, بينما يتمتع المستخدمون العاديون بأنماط استخدام أوسع, بما في ذلك رحلات أطول. وهذا يتماشى مع التوقعات النموذجية للخدمات القائمة على الاشتراك مقارنةً بالاستخدامات غير المنتظمة.

5.8.رسم خرائط لمحطات الدراجات ذات الرحلات الطويلة (أكثر من يوم واحد)

تُحدد الخريطة بدقة التوزيع الجغرافي للمحطات المُستخدمة في رحلات الدراجات الهوائية الطويلة. وتُشير إلى أن المحطات في المناطق المركزية ذات النشاط العالي في واشنطن العاصمة هي الأكثر ارتباطًا بالرحلات التي تستغرق أكثر من يوم. ويشير هذا النمط إلى أن هذه الرحلات الطويلة

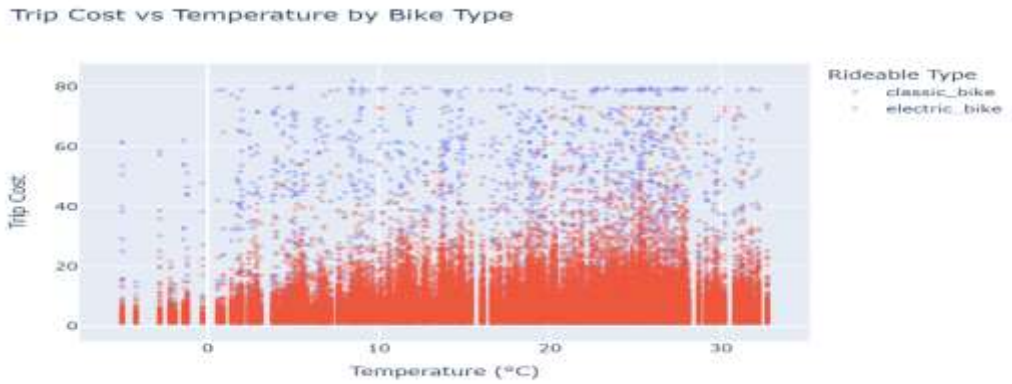
ترتبط في المقام الأول بالاستخدام المكثف ضمن منطقة الخدمة الرئيسية, ربما من قِبل المستخدمين (مثل السياح أو ركاب محددين) الذين يستخدمون الدراجات لفترات طويلة داخل المدينة

5.9. مخطط تكلفة الرحلة مقابل مدتها



يوضح الرسم البياني بوضوح أن معظم الرحلات قصيرة وغير مكلفة. يبدو أن نموذج التسعير يشهد زيادة كبيرة في التكلفة الأولية مع مرور الوقت, إلا أن هذه الزيادة تتضاءل بشكل ملحوظ, أو تستقر التكاليف, في الرحلات الأطول. يُشجع هذا الهيكل على الاستخدام المتكرر قصير الأمد. يشير وجود تكاليف شاذة عالية في الرحلات الطويلة إلى أنه على الرغم من احتمال استقرار متوسط تكلفة التأجير طويل الأمد, إلا أن بعض الرحلات الطويلة جداً قد تُحمّل رسومًا باهظة, ربما بسبب حد أقصى يومي أو غرامة لتجاوز حدود زمنية معينة.

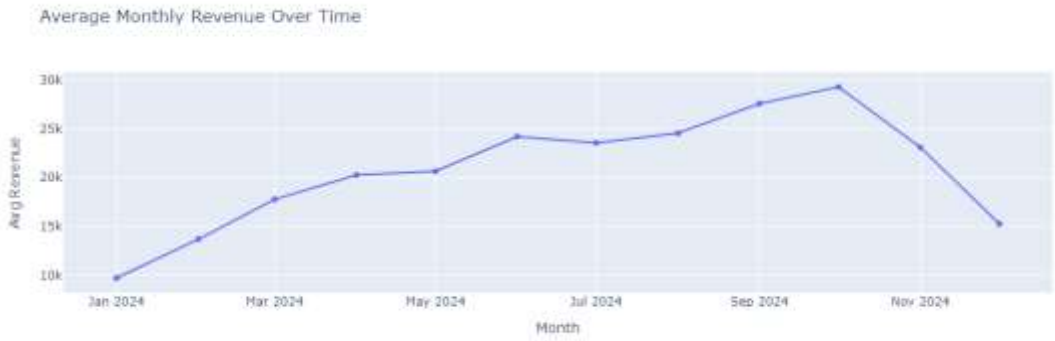
5.10. تكلفة الرحلة مقابل درجة الحرارة حسب نوع الدراجة



يكشف الشكل أنه في حين تُستخدم الدراجات الكلاسيكية والكهربائية بكثافة في نطاق درجة حرارة مريح، فإن الدراجات الكهربائية ترتبط بشكل كبير بالرحلات منخفضة التكلفة. في المقابل، تُمثل الدراجات الكلاسيكية الغالبية العظمى من الرحلات الأعلى تكلفة، مما يشير إلى أنها تُستخدم لفترات أطول أو في سيناريوهات تترتب عليها رسوم أعلى، حتى مع اختلاف درجة الحرارة. لا يبدو أن درجة الحرارة نفسها تُحدد تكلفة الرحلة الفردية بشكل مباشر، بل تؤثر على حجم الاستخدام، مع ملاحظة نطاق أوسع من التكاليف للدراجات الكلاسيكية عبر درجات حرارة التشغيل النموذجية. وهذا يعزز فكرة أن الدراجات الكلاسيكية تُستخدم لأنماط رحلات أكثر تنوعًا (وأحيانًا أطول/أكثر تكلفة) من الدراجات الكهربائية، التي يبدو أن استخدامها أكثر توجهًا نحو رحلات قصيرة وفعالة وبالتالي أرخص.

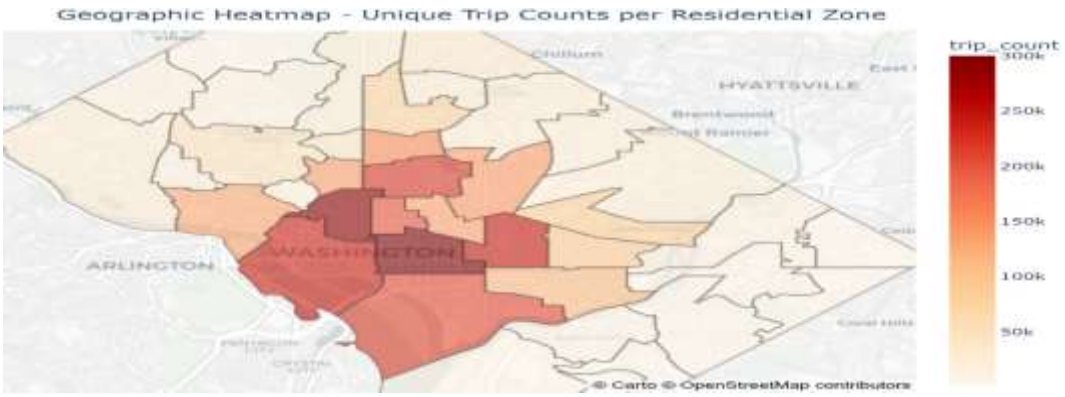
5.11. سلسلة زمنية للإيرادات اليومية والأسبوعية

5.12. سلسلة زمنية للإيرادات الشهرية

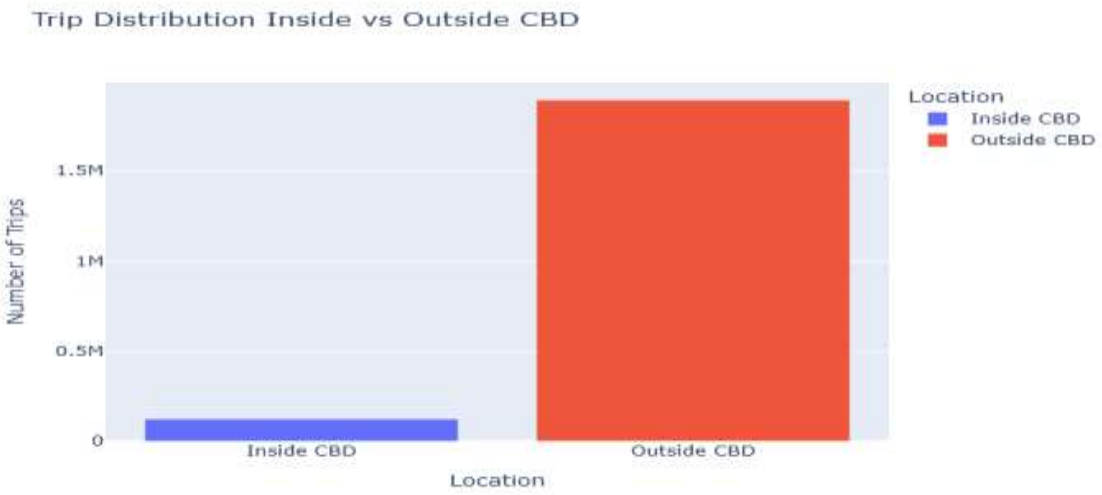


نستنتج من المخططين السابقين تُظهر السلسلة الزمنية للإيرادات بوضوح نمطاً موسمياً قوياً, حيث تبلغ الإيرادات ذروتها خلال الأشهر الأكثر دفئاً (أواخر الصيف/أوائل خريف عام ٢٠٢٤). وتشهد خدمة مشاركة الدراجات انخفاضاً كبيراً في الإيرادات خلال أشهر الشتاء الباردة.

5.8.توزيع الرحلات جغرافيا (داخل المنطقة الرئيسية او في المناطق السكنية)

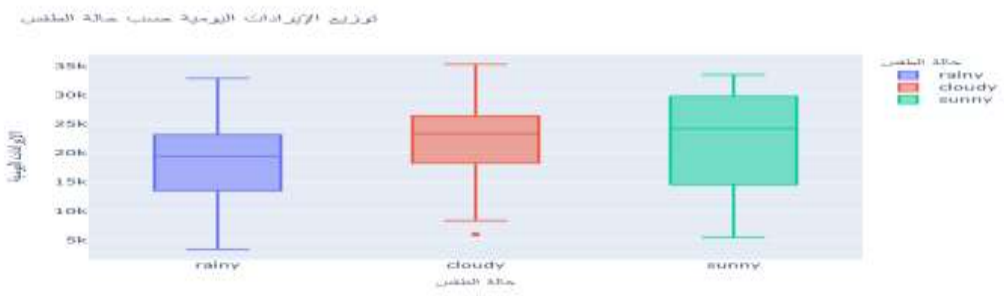


الشكل 14: توزيع الرحلات داخل وخارج منطقة الأعمال المركزية



نستنتج من الشكلين السابقين بوضوح أن نشاط رحلات الدراجات يتركز بشكل كبير في المناطق المركزية المكتظة بالسكان في واشنطن العاصمة, والتي تُعدّ مراكز سكنية وتجارية وسياحية رئيسية.يُقدم الرسم البياني الثاني رؤية مُغايرة للبديهة, ولكنها بالغة الأهمية, عند مُقارنته بالرسم "المسافة إلى منطقة الأعمال المركزية". يظهر الرسم البياني االمسافة الى منطقة الاعمال أن مُعظم الرحلات كانت قريبةً جدًا من منطقة الأعمال المركزية, يُوضح هذا الرسم البياني أن مُجمل الرحلات التي تتم داخل منطقة الأعمال المركزية نادرة جداً.

5.15. الايرادات اليومية حسب حالة الطقس



نستنتج من المخطط بوضوح أن الظروف الجوية تؤثر بشكل كبير على الإيرادات اليومية. يرتبط الطقس "المشمس" بأعلى إيرادات يومية وأكثرها ثباتا، بينما يؤدي الطقس "الممطر" إلى انخفاض كبير في الإيرادات. يقع الطقس "الغائم" بينهما. هذا يعزز الطبيعة الموسمية القوية لنشاط مشاركة الدراجات واعتماده على الطقس، مما يشير إلى أن الطقس الملائم عامل أساسي لزيادة عدد الركاب، وبالتالي زيادة الإيرادات.

التنبؤ بالسلاسل الزمنية :التنبؤ بالاتجاهات المستقبلية

لتوقع الطلب والإيرادات المستقبلية، طبقنا تقنيات التنبؤ بالسلاسل الزمنية.

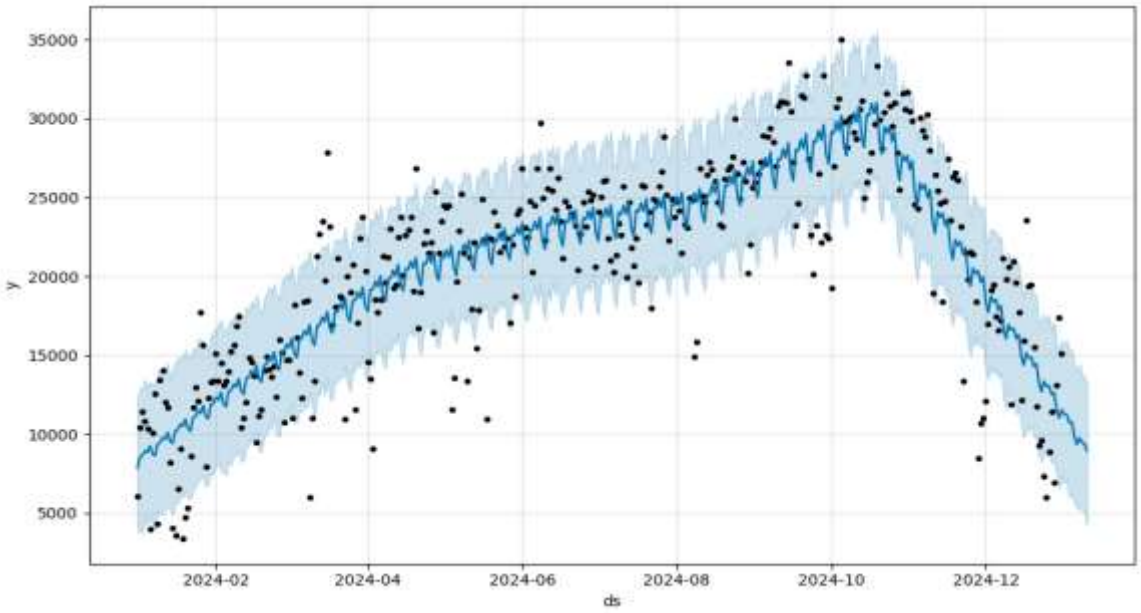
.تجميع الإيرادات اليومية للتنبؤ

بالنسبة لنموذج التنبؤ لدينا، قمنا بتجميع الإيرادات الإجمالية ("trip_total_cost") حسب التاريخ . وقمنا أيضًا بإثراء هذه البيانات اليومية بعمود "weekday" لالتقاط الموسمية الأسبوعية .يسمح هذا النهج المنظم لنموذج التنبؤ بتحديد الأنماط الزمنية والاستفادة منها.

تطبيق نموذجProphet

لقد استخدمنا مكتبةProphet , التي طورتها فيسبوك، لمتانتها في التعامل مع بيانات السلاسل الزمنية ذات المكونات الموسمية القوية والعطلات .تم تدريب النموذج على عمودي) "ds" التاريخ (و) "y" الإيرادات اليومية .(ثم قمنا بإنشاء تواريخ مستقبلية للتنبؤ.

الشكل :16 توقع - Prophet اتجاه الإيرادات اليومية



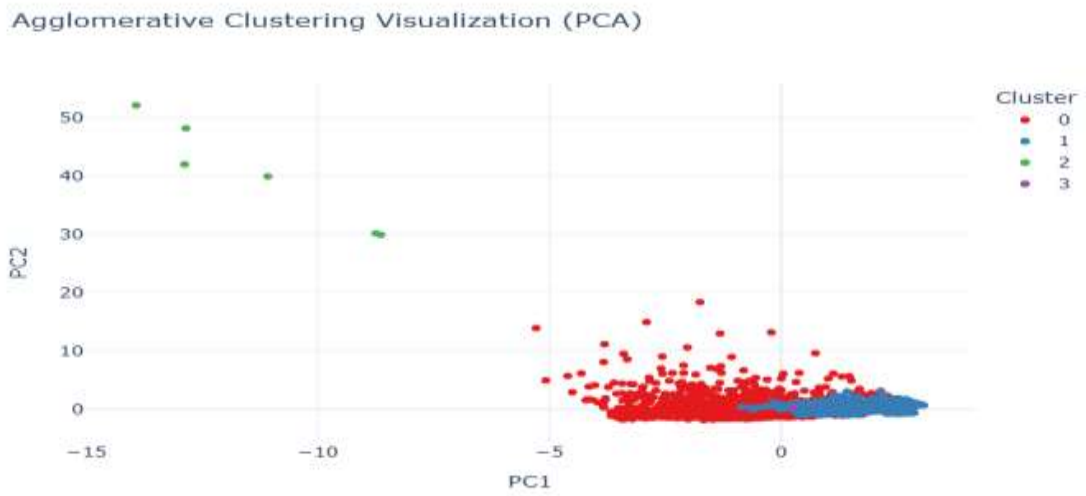
الاستنتاج للشكل :16 يوفر توقع Prophet للإيرادات اليومية إسقاطًا قيمًا للأداء المالي المستقبلي .يتضمن التصور عادةً البيانات التاريخية الفعلية, والاتجاه المتوقع, وفواصل الثقة . يسمح لنا هذا بتوقع فترات الإيرادات المرتفعة والمنخفضة, وفهم الاتجاه الأساسي) النمو أو الانخفاض أو الاستقرار(, وتقييم عدم اليقين المرتبط بالتوقعات .إنها أداة حاسمة لـ التخطيط الاستراتيجي, وتخصيص الموارد, وتحديد التحديات أو الفرص المستقبلية المحتملة.

الشكل :17 توقع - Prophet المكون الأسبوعي

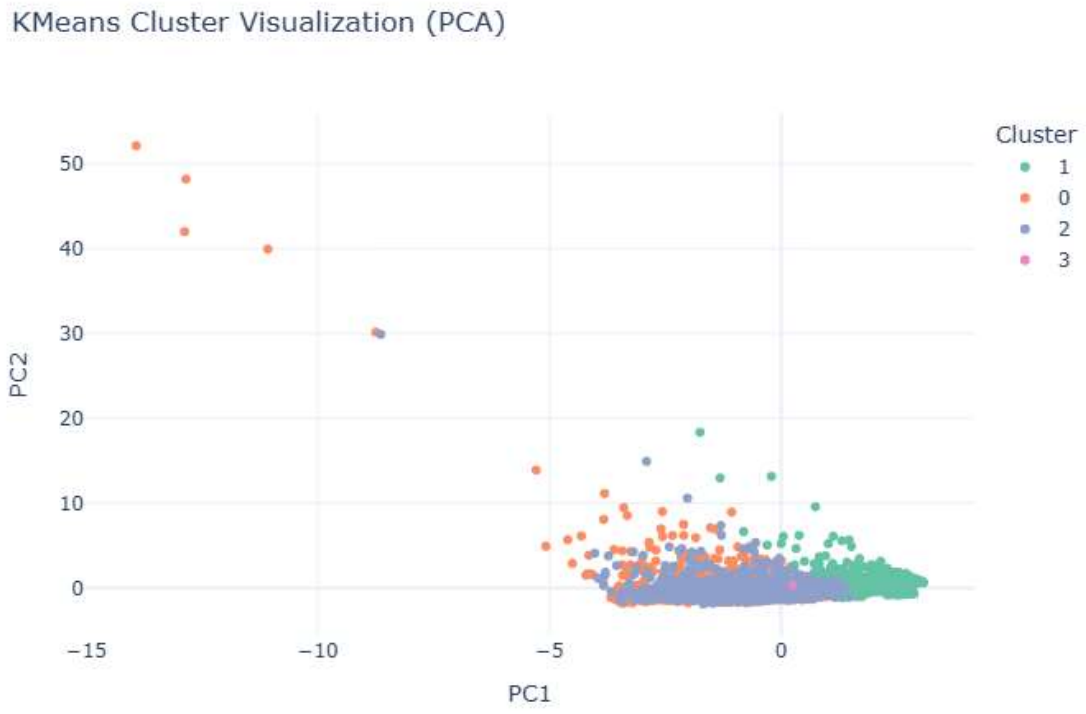
الاستنتاج للشكل :11 يمثل هذا الشكل مكون الموسمية الأسبوعية من نموذجProphet , ويكشف عن أنماط متكررة في إيرادات تأجير الدراجات على أساس أسبوعي .أسبوعيًا :إيرادات منخفضة في بداية الأسبوع) الأحد-الاثنين(, وتصل إلى ذروتها نحو نهاية الأسبوع) الخميس-السبت .(يعد هذا النمط الأسبوعي القوي أمرًا بالغ الأهمية للتخطيط التشغيلي، مما يسمح بإعادة توزيع الدراجات الأمثل، وتوظيف الموظفين، وجهود التسويق لتتوافق مع أيام ذروة الطلب.

7.تحليل تقسيم رحلات العملاء باستخدام clustering

الشكل 18:



الشكل 19:



لفهم أفضل لأنواع الرحلات المختلفة, قمنا بتوظيف تقنيات تعلم الآلة غير الخاضع للإشراف, وتحديداً خوارزميات التجميع العنقودي(Clustering) , لتقسيم بيانات الرحلات إلى مجموعات متميزة. من خلال تطبيق خوارزميتي التجميع التكتلي (Agglomerative Clustering) وK-Meansg, وتصوير النتائج بيانياً باستخدام تحليل المكونات الرئيسية(PCA) , تمكنا من الكشف عن أنماط رئيسية في سلوك الرحلات.

ملاحظات رئيسية من التصورات البيانية

عبر كلا النموذجين, يتضح على الفور وجود هيكل أساسي في البيانات. تُظهر التصورات البيانية تركيزاً كثيفاً للغاية لنقاط البيانات في منطقة واحدة, مع وجود عدد قليل من المجموعات الصغيرة والمتفرقة التي تقع بعيدًا جدًا. يشير هذا الهيكل الأساسي إلى أن الغالبية العظمى من رحلات الدراجات متشابهة جدًا في خصائصها, مما يمثل نمط استخدام "نموذجي". في المقابل, يمثل جزء صغير جدًا من الرحلات قيماً شاذة (outliers) تختلف بشكل كبير عن المعتاد.

نتائج التجميع التكتلي(Agglomerative Clustering)

حدد نموذج التجميع التكتلي أربعة عناقيد, مما سلط الضوء بوضوح على هيكل البيانات هذا:

- العنقودان 0 (الأحمر) و 1 (الأزرق) :هذان العنقودان يشملان المجموعة الضخمة والمركزية من نقاط البيانات. بصريًا, هما غير منفصلين جيدًا ويتداخلان إلى حد كبير, مما يوحي بأن الخوارزمية قامت بتقسيم مجموعة واحدة كبيرة ومتجانسة من الرحلات "القياسية."
- العنقودان 2 (الأخضر) و 3 (الأرجواني) :يحتوي هذان العنقودان على عدد قليل من نقاط البيانات لكل منهما ويقعان بعيدًا عن المجموعة الرئيسية. تمثل هذه النقاط رحلات شاذة للغاية. بالنظر إلى الميزات المستخدمة في التحليل (مثل مدة الرحلة وتكلفتها), من المحتمل أن تكون هذه الرحلات ذات مدد طويلة بشكل استثنائي أو تكاليف عالية بشكل غير عادي أو خصائص نادرة أخرى.

الاستنتاج الرئيسي من هذا النموذج هو فعاليته في الكشف عن القيم الشاذة .فقد نجح في عزل الرحلات الأكثر غرابة, لكنه كان أقل فعالية في إيجاد قطاعات ذات معنى ومتميزة ضمن الكتلة الرئيسية للرحلات الشائعة.

نتائج خوارزميةKMeans

قدم نموذجKMeans , الذي تم إعداده أيضًا لأربعة عناقيد, تقسيمًا أكثر دقة وتفصيلاً:

- تقسيم المجموعة الرئيسية :على عكس التجميع التكتلي, قام KMeans بتقسيم السحابة الكبيرة والمركزية لنقاط البيانات إلى عدة عناقيد متميزة, وإن كانت لا تزال متداخلة إلى حد ما (العناقيد 1 و0 و2). يشير هذا إلى أن KMeans كان قادرًا على تحديد اختلافات أكثر دقة ضمن ملفات تعريف الرحلات "النمطية". يمكن أن تمثل هذه المجموعات الفرعية شخصيات مستخدمين مختلفة, مثل ركاب أيام الأسبوع مقابل راكبي عطلة نهاية الأسبوع الترفيهيين.
- تحديد القيم الشاذة :على غرار النموذج الأول, قام KMeans أيضًا بتجميع القيم الشاذة الأكثر تطرفًا في عنقود واحد منفصل (العنقود 3).

رؤى مقارنة وخاتمة

بمقارنة النموذجين, يمكننا استخلاص الاستنتاجات التالية:

- هيمنة ملف تعريف "الرحلة القياسية :"يؤكد كلا النموذجين أن نوعًا أساسيًا واحدًا من الرحلات يهيمن على استخدام الخدمة. ومن المرجح أن تكون هذه الرحلات قصيرة ومنخفضة التكلفة ويمكن التنبؤ بها.
- وجود قيم شاذة واضحة :هناك مجموعة صغيرة ولكنها متميزة من الرحلات التي تمثل حالات شاذة. قد يكون ذلك بسبب عدم إعادة الدراجات في الوقت المحدد, أو تحركات متعلقة بالصيانة, أو أخطاء في البيانات, ويمكن فصلها بسهولة عن الاستخدام العادي.
- KMeansيقدم تقسيمًا أفضل :لغرض فهم سلوكيات المستخدمين المختلفة داخل المجموعة الأكبر, قدمت خوارزمية KMeans نتيجة أكثر فائدة. فمن خلال تقسيمها للعنقود الكثيف, فإنها توفر نقطة انطلاق لاستكشاف قطاعات أدق داخل قاعدة المستخدمين الرئيسية, وهو ما فشل التجميع التكتلي في تحقيقه.

باختصار، نجح تحليل التجميع العنقودي في تحديد مجموعة أساسية كبيرة من "الرحلات القياسية" ومجموعة صغيرة من "الرحلات الشاذة". وقد أثبت نموذج KMeans أنه أكثر إفادة للأغراض التجارية من خلال الإشارة إلى وجود مجموعات فرعية متميزة حتى ضمن قاعدة المستخدمين القياسية، والتي تستدعي مزيدًا من البحث لتخصيص استراتيجيات التسويق والعمليات والخدمات المقدمة.

الاستنتاج العام :توليف الرؤى لاتخاذ القرارات الاستراتيجية

لقد وفر هذا المسعى الشامل لتعدين البيانات فهمًا متعدد الأوجه لنظام تأجير الدراجات .من الاستكشاف الأولي للبيانات الذي كشف عن أنماط مدة الرحلة وخصائص المستخدمين، إلى هندسة الميزات المعقدة التي أثرت مجموعة بياناتنا بسياق مكاني وزمني، ساهمت كل خطوة في فهم أعمق .أبرز تحليل البيانات الاستكشافي، المدعوم بالتصورات الثاقبة، الاتجاهات الرئيسية في الإيرادات، وسلوك المستخدم، واستخدام المحطات .أخيرًا، مكننا التنبؤ بالسلاسل الزمنية باستخدام Prophetمن توقع الطلب والإيرادات المستقبلية، مما يتيح التخطيط الاستراتيجي الاستباقي .توفر النتائج، مثل انتشار الرحلات القصيرة، وأنماط الاستخدام المميزة للأعضاء مقابل الركاب العاديين، وهيمنة المحطات متوسطة الحجم، ودورات الإيرادات الأسبوعية المتوقعة، معلومات قابلة للتنفيذ لتحسين العمليات، وتعزيز تجربة المستخدم، ودفع النمو المستدام لخدمة مشاركة الدراجات .يمثل هذا التقرير وثيقة حية، شهادة على قوة البيانات في تحويل المعلومات الخام إلى بصيرة استراتيجية.