

الوظيفة الأولى

يوميّات دراجات العاصمة!

الملخص

1. الموضوع: تحليل الرحلات اليومية للدراجات ضمن واشنطن العاصمة
 2. الوحدة: استكشاف البيانات وتحليل السلاسل الزمنية
 3. الأهداف:
- نهدف في هذه الوظيفة إلى تحليل وفهم أنماط استخدام الدراجات في واشنطن العاصمة من خلال دراسة البيانات اليومية للرحلات والعوامل المؤثرة عليها، ويشمل ذلك دراسة تأثير الظروف الجوية، وتحديد مناطق النشاط المرتفع، وتحليل حركة التنقل بين المناطق المختلفة في المدينة.
5. عدد الطلاب: ثلاثة طلاب
 6. الوقت المخصص: عشرون يومًا

قيود التحقيق

1. لغة البرمجة: Python
 2. اطرّات العمل المسموح بها:
- Scipy, Numpy, Pandas/GeoPandas, Plotly, Scikit-Learn, Folium, pygeohash
3. بيئة العمل: Jupyter Notebook, Colab
 4. مجموعات البيانات:
- a. Daily Bikes Trips
 - b. Daily Weather Info
 - c. Stations Info Dataset
 - d. Stations Table
 - e. CBD Polygon
 - f. Metro Bus Stops
 - g. Shuttle Bus Stops
 - h. Residential and Visitor Parking Zones

المتطلبات

1. تنظيف ومكاملة البيانات

a. سد الفجوات:

- لدينا بيانات من مصادر مختلفة ونحتاج إلى دمجها من أجل الحصول على صورة أكثر اكتمالاً لأنماط الاستئجار للدراجات ضمن واشنطن العاصمة، نحن مهتمون بدمج جداول محطات التأجير وحالة الطقس اليومية والمناطق السكنية (اسم المنطقة كما وردت في الجدول عند الانطلاق وعند الانتهاء) مع جدول الرحلات اليومية.
- قم بإجراء عمليات ما قبل المعالجة Preprocessing المناسبة قبل عملية الدمج.
- يجب عليك التأكد من سلامة واكتمال واتساق مجموعة البيانات الموسعة، أي أن تكون على دراية بالمشكلات مثل القيم الفارغة أو الخاطئة وتحولات الأعمدة وتكرار العناصر.

توجيه: يجب عليك استعمال طريقة الدمج المكاني spatial join عند دمج جدول المناطق السكنية، واعتبر أن المسقط (Coordinate Reference System or CRS) الافتراضي للبيانات الجيوغرافية عند التخزين (EPSG:4326) وقم بإجراء التحويلات اللازمة لضمان أن عملية الدمج صحيحة.

b. من الخام إلى المصقول!

- 1 — قم باستخراج سمات التاريخ والوقت البسيطة (العام، الشهر، اليوم، اسم اليوم) وضع كل منها في عمود جديد ضمن جدول الرحلات اليومية.
- 2 — قم ببناء سمة جديدة تعبر عن المدة الإجمالية للرحلة بالدقائق بالاعتماد على تاريخ الانطلاق والانتهاء.
- 3 — قم ببناء سمة جديدة تعبر عن التكلفة الإجمالية للرحلة ضمن الاعتبارات التالية:
 1. في حالة الأعضاء المنتسبين \$3.95 كلفة ثابتة ويضاف لها \$0.05 أو \$0.1 لكل دقيقة في حالة تجاوز مدة الرحلة لـ 45 دقيقة وذلك تبعاً لنوع الدراجة (كلاسيكية، كهربائية) على الترتيب.
 2. في الحالات الأخرى \$1 تكلفة ثابتة ويضاف لها \$0.05 أو \$0.15 لكل دقيقة تبعاً لنوع الدراجة (كلاسيكية، كهربائية) على الترتيب.
 3. تضاف ضريبة \$0.5 عند استعمال محطات المنطقة التجارية الرئيسية
 4. تضاف ضريبة \$3 عند تجاوز يوم كامل
- 4 — قم بإجراء عملية تقطيع لسعة المحطة (كبيرة، متوسطة، صغيرة) معيّنًا حدود كل فئة بطريقة تراها مناسبة
- 5 — قم بحساب المسافة لأقرب موقف حافلة وأقرب موقف مترو عند الانطلاق وعند الانتهاء وضع كل منهم في عمود جديد ضمن جدول الرحلات اليومية.
- 6 — قم بإنشاء سمة ثنائية للدلالة على أن الرحلة كانت تمر (انطلقت أو انتهت) بالمنطقة التجارية الرئيسية للعاصمة.

- قم بحساب المسافة إلى المنطقة التجارية الرئيسية للعاصمة عند الانتهاء من الرحلة. ثم قم بإجراء عملية تقطيع إلى سمة ثنائية للدلالة على أن المحطة بعيدة أو قريبة من المنطقة محددًا العتبة بطريقة تراها مناسبة، وذلك بعد إهمال الرحلات التي انطلقت وانتهت ضمن المنطقة واضعًا قيمة السمة في هذه الحالة بـ null.
 - قم بتجميع المواقع الجغرافية للرحل اليومية ضمن قطاعات جيوغرافية باستخدام تقنية التهشير الجغرافي (Geographic Hashing) وعين دقة مناسبة من وجهة نظرك لعدد البتات المستخدمة في عملية التهشير.
 - ١39 قم بإجراء عملية تجزئة للقطاعات الجغرافية التي حصلت عليها في عملية التهشير الجغرافي إلى 3 فئات تبعًا لمتوسط عدد الرحلات اليومية في كل قطاع (أحمر، أصفر، رمادي) معينًا حدود كل فئة بطريقة تراها مناسبة.
 - قم بإجراء عملية Bining لحالة الطقس إلى ثلاث فئات فقط (مشمس، غائم، ماطر)
 - قم ببناء سلسلة زمنية للإيرادات اليومية الإجمالية (يوم الدفع هو تاريخ الانتهاء) بالاستفادة من جدول الرحلات، يجب تكون السلسلة الناتجة من الشكل العريض أو الطويل (أيهما أسهل بالنسبة لك) بحيث تحوي قيمة الإيرادات وحالة الطقس.
 - قم بإنشاء أي سمة من الممكن أن تساعدك في عملية الاستكشاف أو المكاملة.
- توجيه: قم بإجراء عملية إسقاط للبيانات الجيوغرافية (المضلع أو النقاط) إلى مسقط متري متجانس (EPSG:6933) قبل إجراء أي عملية تتعلق بالمسافة بشكل مباشر أو ضمني، وتذكر أن الأرض ليست كروية الشكل تمامًا وعليه تختلف النتيجة باختلاف المسقط.
- توجيه: يمكنك استخدام النسخة الصغيرة من مجموعة بيانات الرحلات اليومية لتسريع العمل، ويطلب التنفيذ النهائي على المجموعة الأصلية (استخدم Colab في حالة عدم وجود موارد كافية) مبيّنًا الفرق في النتائج إن وجدت.
- تذكرة: إن عملية هندسة السمات قد تجلب مشاكل جديدة ضمن البيانات وعليه يجب عليك مراعاتها بطريقة تراها مناسبة.
- ملاحظة:** يتم تقييم حل الطالب بناءً على **الاستخدام الأفضل** لبيئة العمل **والأسرع** حسابيًا، ويرفض الحل عند وجود خطأ فادح في الاستخدام أو عند إجراء عمليات تساهم في تدمير المعطيات!

2. الاستكشاف والتحليل

a. دراسة حالة الدراجات والمحطات

- قم بتعيين أكثر 5 محطات للانطلاق ثم ارسم مخطط شريطي Bar يعرض الإحصائية
- قم بحساب توزيع عدد الرحلات حسب نوع الدراجة ونوع الاشتراك ثم اعرض النتائج باستخدام مخطط شريطي Bar
- قم برسم مخطط Sunburst يعرض توزيع الرحلات حسب نوع الدراجة ونوع الاشتراك لأكثر 5 محطات للانطلاق
- قم برسم مخطط Histogram يعرض توزيع سعة محطات الدراجات، ثم ارسم مخطط شريطي Bar يعرض توزيع الرحل حسب فئة سعة محطة الانطلاق والانتها.

ماذا تستنتج من كل مخطط؟

b. دراسة مدة الرحلة

- قم برسم مخطط Histogram يعرض توزيع مدة الرحلة بالدقائق محددًا عدد مناسبًا للأشرطة bins
- قم برسم مخطط صندوقي Box يعرض توزيع مدة الرحلة بالدقائق حسب نوع الدراجة
- قم برسم مخطط صندوقي Box يعرض توزيع مدة الرحلة بالدقائق حسب نوع الاشتراك
- قم بتعيين مجموعة الرحلات التي تجاوزت مدتها اليوم الواحد، ثم ارسم محطات الدراجات ضمن خريطة تعرض واشنطن العاصمة محددًا حجم النقطة تبعًا لعدد الرحلات (تجاوزت مدة يوم) التي استقبلتها (مجموع الانطلاق والانتها)

ماذا تستنتج من كل مخطط؟

c. الدراسة المالية

- قم برسم مخطط Histogram ومخطط صندوقي Box ضمن نفس الشكل Figure يعرض توزيع تكلفة الرحلة محددًا عدد مناسبًا للأشرطة bins
- قم برسم مخطط نقطي Scatter لتكلفة الرحلة ضد المدة الزمنية، وقم برسم Line يعرض الاتجاه Trend لهذه النقاط ضمن نفس الشكل Figure
- قم برسم مخطط نقطي Scatter لتكلفة الرحلة ضد درجة الحرارة مميّزًا النقاط حسب فئة الدراجة
- قم بعرض السلسلة الزمنية لإجمالي الإيرادات اليومية والأسبوعية باستخدام مخطط Line وضمن نفس الشكل Figure
- قم بعرض السلسلة الزمنية لمتوسط الإيرادات الشهري ضمن مخطط Line

ماذا تستنتج من كل مخطط؟

d. التحليل المكاني لرحلات الدراجات

- قم بعرض مخطط حراري جيوجرافي Geographic Heatmap لعدد الرحلات في كل منطقة سكنية (الرحلة يمكن أن تنتمي لمنطقتين على الأكثر) وعين إجمالي عدد الرحلات خارج المناطق السكنية.
- قم بعرض مخطط شريطي Bar لتوزيع فئات القطاعات الجغرافية Geographic Hash
- قم بعرض مخطط Histogram لتوزيع البعد عن المنطقة التجارية الرئيسية والمسافة لأقرب نقطة مترو والمسافة لأقرب موقف حافلة محددًا عددًا مناسبًا للBins
- قم بعرض مخطط شريطي Bar لتوزيع الرحلات فيما إذا كانت الرحلة داخل المنطقة التجارية الرئيسية بالكامل (انطلقت وانتهت فيها) أم خارجها
- قم بعرض مخطط شريطي Bar لتوزيع الرحلات التي مرت ضمن المنطقة التجارية الرئيسية حسب نوع الدراجة ونوع الاشتراك
- قم بدراسة الارتباط بين فئة البعد عن المنطقة التجارية الرئيسية ونوع الاشتراك باستخدام تحليل chi square

ماذا تستنتج من كل مخطط؟

e. دراسة تأثير حالة الطقس

- قم بعرض متوسط درجة الحرارة وسرعة الرياح والرطوبة اليومية باستخدام مخطط Line وضمن نفس الشكل Figure
- قم بعرض مخطط Box للإيرادات اليومية حسب حالة الطقس (مشمس، غائم، ماطر)
- قم بدراسة الارتباط **الخطي** بين درجة الحرارة والرطوبة والإيرادات اليومية
- قم بدراسة الارتباط بين حالة الطقس (مشمس، غائم، ماطر) ونوع الدراجة باستخدام تحليل chi square

ماذا تستنتج من كل مخطط؟

توجيه: إن وجود عناصر **شاذة** ضمن المعطيات قد يؤثر **سلبيًا** على المخططات المرسومة، وعليه تحرى الدقة أثناء الرسم وانتبه إلى نتائج الإحصائيات التي تقوم بها.

توجيه: عند رسم مخطط يتسبب في إنهاء الجلسة وإعادة تشغيل النواة (Kernel) أو عدم استجابة Colab لعمليات الحفظ أو وجود بطء في صفحة الويب قم بالتخلص من ميزة التفاعلية J Plotly وارسم المخطط كصورة ثابتة وراجع المدرس في حال واجهتك أي مشكلة أثناء الرسم.

ملاحظة: **يهمل** كل مخطط لا يتم شرحه من قبل الطالب ضمن المفكرة ويترتب على ذلك خصم علامة الطلبات المتعلقة بالمخطط **كليًا أو جزئيًا** حسب ما يراه المدرس مناسبًا.

3. اصطيات الأنماط.

a. التحليل الزمني

- قم بالتأكد من أن البيانات التي حصلت عليها هي بالفعل سلسلة زمنية، وقم بإجراء الإصلاحات اللازمة في حال وجود أي مشكلة.
- قم ببناء baseline للتنبؤ بالإيرادات المستقبلية على مدى 10 أيام بالطريقة التي تراها مناسبة.
- قم بنمذجة مكونات السلاسل الزمنية باستعمال Prophet واستخدم النموذج للتنبؤ بالمستقبل وقم بعمليات التوليف tuning لبناء أفضل نموذج.
- قارن بين النماذج باستعمال طرق التقييم المناسبة.
- ماذا تستنتج من كل مكون؟

b. التحليل العام لأنماط الاستخدام

- قم بإجراء عملية Sampling للبيانات إلى درجة تستطيع تحليلها ومعالجتها بحيث تكون العينات **شاملة** قدر الإمكان وقم بتحديد عدد كافي للسعات بالطريقة التي تراها مناسبة.
- قم بإجراء عملية Clustering باستخدام تقنيات التعلم التلقائي (**ثلاث خوارزميات** على الأكثر **واثنين** على الأقل) ثم حاول وصف العناقيد وقارن بينها.
- قم بإجراء عمليات ما بعد المعالجة المناسبة ثم حدد أهم السمات التي أثرت على قرار أفضل نموذج دربته لكل خوارزمية استخدمتها مبيئاً رأيك

4. التوثيق

كما تعلم، يعد تقرير التوثيق مكوناً حاسماً في أي مشروع أو مهمة، لأنه بمثابة سجل شامل لما تم إنجازه وكيف تم تحقيقه، مما يسمح لأعضاء الفريق وأصحاب المصلحة في المستقبل بفهم أهداف المشروع والجدول الزمني والتحديات والحلول.

ستكون مسؤولاً عن إنشاء تقرير باللغة العربية يلخص ما قمت به وما توصلت إليه، وتذكر أن جودة التقرير المكتوب لا يفيدك أنت وفريقك فحسب، بل يخدم أيضاً كأصل قيم للمقارنة بين عملك وعمل أقرانك، وعليه تحرى الدقة والترتيب في كتابة تقرير التوثيق.

خذ وقتك وكن دقيقاً في وصفك وتأكد من تضمين جميع المعلومات ذات الصلة وقدم تفسيرات وأمثلة واضحة لدعم نتائجك واستخدم الأدوات المناسبة لتصدير صور المخططات ونسق النتائج ضمن جدول إن لزم الأمر ولا تضع أي لقطة شاشة أو جزء من الكود ضمن التقرير!

5. مهمة سرية!

قد يبدو التنقيب عن المعارف أمرًا شاقًا في البداية، ولكن تذكر أن كل جزء من البيانات يحمل رؤى قيمة في انتظار الكشف عنها، ومن خلال مهاراتك التحليلية وتصميمك تستطيع الغوص في أعماق البيانات واستكشاف تعقيداتها واستخراج أنماط وحقائق ذات مغزى، وفي حين أنه من المهم الالتزام بمتطلبات الوظيفة، لا تدعها تحد من رحلتك الاستكشافية. اسمح لنفسك بالتفكير خارج الصندوق والتفكير في زوايا ووجهات نظر مختلفة.

يمكنك استخدام أي جدول من قاعدة البيانات حسب حاجتك، ورسم أي مخطط أو حساب أي احصائية لدعم استنتاجاتك، وتذكر أن التنقيب هو عملية ديناميكية وتكرارية، وعليه، تأكد من كتابة كود نظيف وقابل للقراءة وإعادة الاستخدام وحاول تنظيم المفكرة قدر الإمكان.

تذكر أن pipes ضمن مكتبة Pandas واستخدم أكبر عدد ممكن من وظائف المكتبة بدلاً من إعادة اختراع العجلة!

ابق فضوليًا وكن مبدعًا ودع البيانات ترشدك في رحلتك، واعلم أن مهارة رواية القصص تعد من أكثر المهارات طلبًا إن جمعت مع علم البيانات! إذ أن طريقة العرض للجمهور المتلقي أهم بكثير من النتائج.

قيود وقواعد تنظيمية

1. يعتبر طلب [التحليل العام لأنماط الاستخدام](#) طلبًا إضافيًا لتعويض **النقص** في درجات تقييم المخاطر. وتقدر العلامة وفقًا لحجم الجهد المبذول والمعلومات المستخرجة من البيانات.
2. تعتبر **المهمة السرية** طلبًا إضافيًا وتقسم الدرجة فيه إلى قسم لتعويض **النقص** في درجات تقييم المخاطر وقسم لجبر أخطاء الطالب ضمن الوظيفة وفقًا لحجم الجهد المبذول وحد أعلى يحدد لاحقًا.
3. الوظيفة تتطلب مجموعة مكونة من 3 طلاب لا أقل ولا أكثر *ولا مجال* للجدال في العدد المسموح. تحتاج الوظيفة لتنتهي في الوقت المحدد تعاون الأفراد الثلاثة، كما أن علامة الطلاب في المجموعة الواحدة ليست موحدة. يحصل الطلاب على جزء من العلامة على حل الوظيفة والجزء الآخر على مقدار عمل الطالب بها، بحيث يقدر عمل كل طالب من خلال مقابلات في نهاية الفصل يحدد موعدها لاحقًا.
4. تأكد قبل تسليم الملف أن جميع الخلايا منفذة بشكل كامل والنتائج معروضة فيه. وتأكد أن الملف يعمل وأنه قابل للقراءة بوضوح **وبدون أي مخطط تفاعلي** (جميع المخططات ترسم كصور ثابتة قبل التسليم)، قم بفتح ملف الـ html وتأكد أنه صحيح وأن حجم الملف المسلم صغير. في حال كان كبيراً تأكد من أنك لم تقم بطباعة كل البيانات بشكل غير مقصود فيه، في حال تجاوز حجمه الحد المسموح (10M) لن يتم رفعه.
5. سيتم الإعلان عن نموذج التقديم في موعد لاحق.
6. سيتم معاقبة الوظائف المتأخرة على النحو التالي:
 - a. تأخير يوم واحد: خصم 15% من العلامة
 - b. تأخير يومين: خصم 30% من العلامة
 - c. ثلاث أيام متأخرة أو أكثر: لن يتم قبول الوظيفة وتحصل المجموعة على درجة الصفر كاملة!
7. خصص وقتًا كافيًا لإكمال الوظيفة دون تسرع أو طرق ملتوية، ابدأ مبكرًا وخطط مسبقًا للتأكد من أن لديك الوقت الكافي للحصول على البيانات والمعالجة المسبقة والتحليل، ولا تنتظر حتى اللحظة الأخيرة فقد يؤدي ذلك إلى تسريع العمل وزيادة احتمالية سوء السلوك الأكاديمي.
8. يجب أن يكون كل العمل خاص بك. لا تنسخ أو تعيد صياغة أعمال الآخرين دون **اقتباس** مناسب لتوضيح **المصدر**. إذ أن الغش لا يؤدي إلى تقويض عملية التعلم فحسب، بل ينتهك أيضًا الثقة بين الطالب والمدرس ويمكن أن يؤدي إلى عواقب وخيمة، لذلك تجنب مشاركة عملك مع الآخرين أو السماح للآخرين باستخدام عملك.
9. لا تستخدم أي مادة أو مكتبة غير مصرح بها ما لم يسمح المدرس صراحة بذلك (يمنع استخدام matplotlib, seaborn على سبيل المثال لا الحصر) ويستثنى من هذه القاعدة المخططات الحرارية Heatmap أو الخرائط الجغرافية ويسمح باستخدام أي أداة **برمجية** تسهل عملية رسمها.
10. افهم الفرق بين التعاون والغش، يُسمح بالتعاون طالما أنه لا يتضمن جزءًا من عمل زملائك.
11. كن صريحًا وشفافًا بشأن عملك.
12. إذا كانت لديك أسئلة بشأن الوظيفة فاطلب التوجيه من المدرس.
13. تأكد من الاطلاع على دليل مكتبة Pandas أو Plotly حتى لا تقع في أخطاء فادحة لسوء الاستخدام، ولا تحصر نفسك بما تم اعطائه ضمن المحاضرة!
14. لا تترك أي رسالة خطأ أو رسالة طباعة طويلة جدًا ضمن المفكرة!

15. عند مراجعة كود الوظيفة في حال تبين أنه مأخوذ بالكامل من الانترنت سوف تنال الوظيفة درجة الصفر كاملة!
16. لا يجب عليك التكلف في استنتاجاتك، اكتبها بأي طريقة تناسبك وبأي لغة تريدها (العربية أو الإنجليزية)، موضحة المصطلحات إن وجدت باللغة الإنجليزية وذلك ضمن قوسين.
17. في حالة عدم اقتباس مصدر الكود هذا يعني أنك صاحب الفكرة، وهنا انتبه إلى العبء الممكن أن تتحمله في حال تبين خلاف ذلك!، وتذكر أن أي مخالفة للضوابط والقواعد المذكورة وسوء الممارسة الأكاديمية يستلزم العقوبة وفق ما يراه المدرس مناسباً.

وتذكر:

تنفذ الإرشادات وتعليمات التسليم حرفياً دون إبداعات إضافية، لا زيادة ولا نقصان.
عند وجود أي تشابه بين وظيفتي مجموعتين ستخسر المجموعتان العلامة معاً دون مراجعة أي منهما

