

الوظيفة الأولى

فوضى المدينة

الملخص

1. الموضوع: استكشاف حوادث السيارات في المتاهة الحضرية لمدينة شيكاغو
2. الوحدة: استكشاف البيانات
3. الأهداف:

مهمتنا هي إجراء تحليل استكشافي لحوادث السيارات في مدينة شيكاغو باستخدام مجموعة بيانات شاملة تتكون من ستة مكونات رئيسية: حوادث السيارات، الأفراد المتورطين، المركبات، حدود مناطق الشرطة، حدود المنطقة التجارية المركزية وحدود شوارع المدينة، باستخدام قوة البيانات المخفية نهدف إلى الحصول على رؤى قيمة حول الأنماط والعوامل المؤثرة في حوادث السيارات داخل المدينة.

من خلال هذا التحليل، نسعى لتسليط الضوء على أهمية السلامة العامة وتخطيط الطرق المرورية وتنظيم القوانين، فمن خلال فهم الشبكة المعقدة لحوادث السيارات يمكننا تحديد المناطق ذات المخاطر العالية والعوامل المشتركة التي تؤدي إلى الحوادث. وبهذه المعرفة يمكن توجيه الجهود نحو تحسين إجراءات سلامة الطرق وتخطيط البنية التحتية وإدارة حركة المرور لخلق بيئة نقل أكثر أماناً وانسجاقاً.

5. عدد الطلاب: ثلاثة
6. الوقت المخصص: 14 يوم (حتى تاريخ 2024-5-26 الساعة 11:59 مساءً)

قيود التحقيق

1. بيئة العمل: Jupyter Notebook, Colab
2. مجموعات البيانات:
Traffic Crashes, Streets Center Lines, Police Beats Boundaries, CBD Boundary

المتطلبات

1. تنظيف ومكاملة البيانات

a. سد الفجوات:

- لدينا بيانات من مصادر مختلفة ونحتاج إلى دمجها حتى نحصل على صورة أكثر اكتمالاً لحوادث السيارات ضمن مدينة شيكاغو، وفي حين أن جداول البيانات تحوي الكثير من الأعمدة (أكثر من 20 عمود لكل جدول). انتبه إلى أنها قد لا تحوي معلومات كافية. قم بتحديد الأعمدة التي تحتاجها موضحاً سبب اهمالك لباقي الأعمدة في كل جدول من قاعدة البيانات.
- قبل البدء بعملية التحليل يجب عليك تنظيف مجموعة البيانات أي يجب عليك حل مشكلات القيم الفارغة والمتطرفة وكذلك المتناقضة بطريقة تراها مناسبة.

b. من الخام إلى المصقول!

- i. قم باستخراج العام الذي حدث به الحادث المروري وضعه ضمن عمود جديد بشكل مماثل لعمود الشهر والساعة.
- ii. استنفد من جدول الأفراد المتورطين ضمن الحادث وقم بحساب عدد الركاب الموجودين ضمن كل مركبة (بخلاف السائق) ثم احسب متوسط عدد الركاب ومتوسط عمر الراكب في كل حادث مروري ضمن قاعدة المعطيات.
- iii. استنفد من جدول المركبات لحساب عدد المركبات المتورطة في الحادث المروري.
- iv. استنفد من جدول المركبات وقم بتحديد صنف عمر المركبة (عتيق، قديم، جديد) وذلك بالاستفادة من عمود تاريخ الصنع وتاريخ الحادث معيّنًا حدًا مناسبًا من وجهة نظرك للتصنيف، ثم أنشئ صنفًا متعددًا (Multi Categorical) لأصناف عمر المركبات المتورطة في الحادث المروري على شكل قائمة comma separated حيث تحوي القائمة الأصناف المميزة فقط (Unique) بغض النظر عن تكرار كل صنف.
- v. قم بإجراء عملية Binning لقائمة أصناف عمر المركبات على النحو التالي (قديم، جديد، مختلط) وحولها إلى سمة اسمية تقليدية (Categorical).
- vi. قم بتجميع المواقع الجغرافية للحوادث المرورية ضمن قطاعات جيوغرافية باستخدام تقنية التهشير الجيوغرافي (Geographic Hashing) وعين دقة مناسبة من وجهة نظرك لعدد البتات المستخدمة في عملية التهشير.
- vii. قم بإجراء عملية تجزئة للقطاعات الجغرافية التي حصلت عليها في عملية التهشير الجيوغرافي إلى 3 فئات تبعًا لعدد الحوادث المرورية في كل قطاع (أحمر، أصفر، رمادي) معيّنًا حدود كل فئة بطريقة تراها مناسبة.
- viii. قم بحساب طول كل شارع ضمن مدينة شيكاغو بالكيلومتر لتقوم بعدها بإجراء عملية تقطيع (Discretization) إلى ثلاث فئات (قصير، متوسط، طويل) معيّنًا حدود كل فئة بطريقة تراها مناسبة ثم ادمج النتيجة التي حصلت عليها مع جدول الحوادث المرورية.

- ix. قم باستخدام المضلع الجيوغرافي (Geographic Polygon) الذي يحدد المنطقة التجارية المركزية ضمن المدينة من أجل حساب بعد كل حادث عن هذه المنطقة بالكيلومتر، ثم قم بإجراء عملية تقطيع إلى ثلاث فئات (قريب، متوسط البعد، بعيد).
 - x. قم بإجراء عملية Binarization لعمود البعد عن المنطقة التجارية لتحديد فيما إذا كان الحادث ضمن المنطقة أم خارجها.
 - xi. استفد من جدول حدود المقاطعات الشرطية وحدد المقاطعة الشرطية التي حدث بها كل حادث مروري.
 - xii. قم بإنشاء أي سمة من الممكن أن تساعدك في عملية الاستكشاف أو المكاملة.
- توجيه: قم بإجراء عملية إسقاط للمضلع الجيوغرافي إلى مسقط ذو مساحة مترية متساوية -متجانس- (EPSG:6933) قبل حساب بعد النقطة الجيوغرافية عن ذلك المضلع، وتذكر أن الأرض ليست كروية الشكل تمامًا وعليه تختلف النتيجة باختلاف المسقط.
- تذكرة: إن عملية هندسة السمات قد تجلب مشاكل جديدة وعليه يجب عليك حلها والانتباه إلى الحالات المختلفة وما الفائدة الحقيقية من كل بسمة حتى لا تتسبب في ضياع المعلومات.

2. الاستكشاف والتحليل

a. دراسة أسباب الحوادث المرورية

- i. قم برسم مخطط Bar لعدد الحوادث حسب السبب الرئيسي بتقدير ضابط الشرطة (PRIM_CONTRIBUTORY_CAUSE).
 - ii. قم برسم مخطط Bar لعدد الحوادث حسب ظروف الإضاءة (LIGHTIING_CONDITION) وفئة الضرر المقدّر من قبل ضابط الشرطة (DAMAGE).
 - iii. قم برسم مخطط Bar لعدد الحوادث حسب تصنيف خطورة الحادث (CRASH_TYPE) ونوع الاصطدام الأول (FIRST_CRASH_TYPE).
 - iv. قم برسم مخطط Bar لعدد الحوادث حسب نوع الطريق (TRAFFICWAY_TYPE) وظروف الإضاءة (LIGHTIING_CONDITION).
 - v. حاول شرح النتائج بالاستفادة من معرفتك بالمجال.
- ### b. دراسة تاريخ الحوادث المرورية
- i. قم برسم مخطط Bar لعدد الحوادث المرورية حسب الساعة (CRASH_HOUR) وفئة الضرر المقدّر من قبل ضابط الشرطة (DAMAGE).
 - ii. قم برسم مخطط Line لعدد الحوادث الكلي في الشهر من كل سنة (TOTAL COUNT PER MONTH PER YEAR).
 - iii. قم برسم مخطط Box لتوزيع عدد الحوادث الشهرية خلال السنة.
 - iv. قم برسم مخطط Sunburst لعدد الحوادث حسب يوم الأسبوع (CRASH_DAY_OF_WEEK) وفئة الضرر المقدّر من قبل ضابط الشرطة (DAMAGE).
 - v. حاول شرح النتائج بالاستفادة من معرفتك بالمجال.

- c. التحليل المكاني للحوادث المرورية
- قم برسم مخطط Bar لعدد الحوادث المرورية في كل قطاع جيوغرافي
 - قم برسم مخطط Box للمسافة عن المنطقة التجارية المركزية حسب فئة الضرر المقدر من قبل ضابط الشرطة (DAMAGE).
 - قم برسم مخطط Box لطول الشارع ضمن المدينة.
 - قم برسم مخطط Bar لعدد الحوادث المرورية بحسب فئة طول الشارع.
 - قم برسم مخطط Bar لعدد الحوادث المرورية بحسب فئة البعد عن المنطقة التجارية المركزية.
 - هل هناك ارتباط بين فئة طول الشارع وفئة البعد عن المنطقة التجارية المركزية (Chi Square Analysis)
 - حاول شرح النتائج بالاستفادة من معرفتك بالمجال.

- d. دراسة حالة الركاب
- قم برسم مخطط Scatter حيث تمثل كل نقطة (عمر السائق، جنس السائق، عمر السيارة بتاريخ الحادث المروري).
 - هل هناك ارتباط بين فئة عمر السيارة وفئة عمر السائق (Chi Square Analysis)؟
 - قم برسم مخطط Box لمتوسط عمر الركاب باستثناء السائق (أو شرح جدول التوصيف الخاص بالسمة)
 - قم برسم مخطط Box لعدد الركاب باستثناء السائق (أو شرح جدول التوصيف الخاص بالسمة)
 - قم برسم مخطط Bar لفئة عمر السيارة
 - حاول شرح النتائج بالاستفادة من معرفتك بالمجال.
- e. دراسة الحالات المتطرفة

- قم بتعيين مجموعة الحوادث المرورية المتطرفة أو الغريبة بطريقة تراها مناسبة.
 - حاول أن تستكشف أسباب تطرف هذه الحالات مبيناً رأيك.
- توجيه: يمكنك استخدام Cramer's V لقياس قوة الارتباط إن وجدت.

(انظر [المصادر](#) لمزيد من المعلومات)

توجيه: عند رسم مخطط يتسبب في إنهاء الجلسة وإعادة تشغيل ال kernel أو عدم استجابة Colab لعمليات الحفظ قم بالتخلص من ميزة التفاعلية ل 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 كلياً أو جزئياً حسب ما يراه المدرس مناسباً.

3. مهارة البحث

قم بالبحث عن ورقة بحثية تدرس طريقة معالجة القيم المحذوفة المقنعة disguised missing values. افهم الطريقة ولخصها بما لا يتجاوز الصفحتين مع مراعاة كون البحث حديث قدر الإمكان (من 2020)، واكتب في التلخيص ما يلي:

- اسم المجلة الناشرة
- شرح طريقة الحل
- مجموعات البيانات التي استخدمت للاختبار ونتائج الاختبار
- سلبيات أو قيود الطريقة

4. 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، وإعادة تحقيق أي متطلب أو إجرائية وتذكر أن التنقيب هو عملية ديناميكية وتكرارية، وعليه، تأكد من كتابة كود نظيف وقابل للقراءة وإعادة الاستخدام وحاول تنظيم المفكرة قدر الإمكان،

اطلع على وظيفة الـ pipes ضمن مكتبة Pandas واستخدم أكبر عدد ممكن من وظائف المكتبة بدلًا 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

مرفق مع الوظيفة:

- ملف notebook، يطلب منك أن تقوم بملئ الخلايا وفق الطلبات المدرجة أعلاه. في حال طلب منك تثبيت ملاحظاتك قم بإضافة خلية نصية واكتب ملاحظاتك بلغة سليمة. لا تنسى أن تضع شرح بسيط قبل التوابع في حال كانت معقدة وتستلزم الشرح وحافظ على نظافة الكود.
 - ملف ورد، تكتب فيه نتائج الطلب الثالث دون تغيير في شكل الملف.
1. آخر موعد لتسليم الوظيفة يوم الأحد 2024-5-26 الساعة 11:59 مساءً، سيتم معاقبة الوظائف المتأخرة بخصم 5% من العلامة عن كل يوم تأخير.
 2. الوظيفة تتطلب مجموعة مكونة من 3 طلاب لا أقل ولا أكثر ولا مجال للجدال في العدد المسموح، وغير مسموح بتغيير أفراد المجموعة في الوظائف القادمة، تحتاج الوظيفة لتنتهي في الوقت المحدد

- تعاون الأفراد الثلاثة (الوقت محسوب بدقة)، 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، معرفة عمل كل طالب ومقداره سيجري من خلال مقابلات في نهاية الفصل يحدد موعدها لاحقاً.
3. قم بتغيير اسماء الملف بكتابة اسم أعضاء الفريق باللغة العربية مكان [your_name]، قم برفع نسختين من ال notebook الأول بصيغة ipynb والثاني بصيغة html بالإضافة إلى ملف pdf للطلب الثالث على الرابط الآتي دون ضغطه:
<https://forms.gle/naYmp8RLUnTdwUCZA>
4. تأكد قبل تسليم الملف أن جميع الخلايا منفذة بشكل كامل والنتائج معروضة فيه. وتأكد أن الملف يعمل وأنه قابل للقراءة بوضوح، قم بفتح ملف ال 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. الجزء الأكبر من علامة الوظيفة تعطى بناء على استنتاجاتك لا الكود فقط.
6. خصص وقتاً كافياً لإكمال الوظيفة دون تسرع أو طرق ملتوية، ابدأ مبكراً وخطط مسبقاً 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.
7. 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 أيضاً الثقة بين الطالب والمدرس ويمكن أن يؤدي إلى عواقب وخيمة، لذلك تجنب مشاركة عملك مع الآخرين أو السماح للآخرين باستخدام عملك.
8. افهم الفرق بين التعاون والغش، يُسمح بالتعاون طالما أنه لا يتضمن جزءاً من عمل زملائك.
9. كن صريحاً وشفافاً بشأن عملك.
10. إذا كانت لديك أسئلة بشأن الوظيفة فاطلب التوجيه من المدرس.
11. قبل البدء بتحقيق الوظيفة اقرأ كل طلب بتمعن وكذلك انظر إلى كتيب قاعدة البيانات لتحصل على توصيف دقيق لكل عمود ضمن الجداول.
12. 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!
13. قبل تسليم الوظيفة قم برسم جميع المخططات المطلوبة كصور ثابتة غير تفاعلية وتأكد أن ملف ال HTML يحوي الخرج الصحيح لكل خلية.
14. لا تترك أي رسالة خطأ أو رسالة طباعة طويلة جداً 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)، موضحاً 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 أن:

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

مدرسو المادة: علا طبال، عدنان قطان، رفيف زينية

المصادر:

1. [Chicago Crashes Dataset](#)
2. [DM 2024](#)
3. [Pandas Docs](#)
4. [Plotly Docs](#)
5. [Geohash Wiki](#)
6. [How to easily check for associations between categoricals](#)

