الوظيفة الأولى
تحدي كأس العالم

الملخص

1. الموضوع: تحليل أحداث كأس العالم
2. الوحدة: استكشاف البيانات
3. الأهداف:

مسابقة كأس العالم لكرة القدم، وغالباً ما يطلق عليها مسابقة كأس العالم، وهي مسابقة كرة قدم دولية تتنافس عليها الفرق الوطنية للرجال من أعضاء الاتحاد الدولي لكرة القدم (FIFA) وهي الهيئة الحاكمة للرياضة بشكل عالمي.

اعتبارًا من كأس العالم لكرة القدم 2022 ، أقيمت 22 بطولة وتنافس ما مجموعه 80 فريقًا وطنيًا. وفازت بالكأس ثمانية منتخبات وطنية.

نظرًا لأن كأس العالم هي أكثر بطولات كرة القدم شهرة في العالم، فضلاً عن أنها الحدث الرياضي الأكثر مشاهدةً ومتابعةً على نطاق واسع، فإن مهمتك هي استخراج حقائق مثيرة للاهتمام حول البطولات والمباريات والفرق واللاعبين والملاعب.

5. عدد الطلاب: اثنان
6. الوقت المخصص: ثمانية أيام

قيود التحقيق

1. لغة البرمجة: Python
2. اطارات العمل المسموح بها: Scipy, Numpy, Pandas, Plotly, Scikit-Learn
3. بيئة العمل: Jupyter Notebook, Colab
4. مجموعات البيانات:

The Fjelstul World Cup Database, World Cup Attendance Dataset, World Cup Referees Dataset

المتطلبات

1. تنظيف ومكاملة البيانات

a. سد الفجوات:

- لدينا بيانات من مصادر مختلفة ونحتاج إلى دمجها من أجل الحصول على صورة أكثر اكتمالاً لأحداث كأس العالم، نحن مهتمون بدمج مجموعة بيانات الحضور وقاعدة بيانات Fjelstul، وتُترك مجموعة بيانات الحكام للطالب كتدريب إضافي، يجب أن تحتوي مجموعة البيانات النهائية على جميع السمات من جدول المباريات وكذلك سعة الملعب من جدول الملاعب في قاعدة Fjelstul، وعدد الحضور الجماهيري من مجموعة بيانات الحضور.
- يمكنك استخدام أيٍ من جداول قاعدة بيانات Fjelstul حسب الحاجة في عملية المكاملة.
- يجب عليك التأكد من سلامة واكتمال مجموعة البيانات الموسعة، أي أن تكون على دراية بالمشكلات مثل القيم الفارغة وتحولات الأعمدة وتكرار العناصر

توجيه: ركز على match_date ، away_team_code ، home_team_code وانتبه على فهارس كل جدول لتسهيل عملية الدمج.

b. من الخام إلى المصقول!

- وفقًا لقواعد الفيفا، يُسمح للاعب بتمثيل منتخب واحد فقط في المسابقات الرسمية، بما في ذلك كأس العالم، ومع ذلك كانت هناك حالات قليلة لعب فيها اللاعب لأكثر من فريق وطني واحد في مسيرته، ولكن ليس في نفس البطولة. يطلب إنشاء مجموعة بيانات players_teams باستخدام جدول الفرق (Squads) واللاعبين (Players) في قاعدة بيانات Fjelstul.
- يجب أن تتضمن مجموعة البيانات الناتجة معرف اللاعب، واسمه، واسم العائلة، وعدد البطولات، وقائمة البطولات التي شارك بها من جدول اللاعبين إلى جانب قائمة أسماء الفرق، وقائمة رموز الفرق، وعدد الفرق التي مثلها اللاعب خلال مسيرته في كأس العالم.
- يجب أن يكون تنسيق القوائم الجديدة هو نفسه التنسيق القديم، أي أنه يماثل تنسيق قائمة البطولات التي شارك بها.

2. هندسة السمات

- إنشاء السمة total_goals_in_match (فقط في جدول المباريات).
- إنشاء سمة ثنائية match_for_host للدلالة على أن المباراة يلعب بها مضيف البطولة بغض النظر عن الجانب الذي يلعب به (فقط في جدول المباريات).
- إنشاء سمة used_capacity_ratio. (فقط في جدول المباريات).
- استخدم التقطيع (Discretization) لإنشاء attendance_category وكذلك relative_attendance_category اعتمادًا على عدد الحضور ونسبة السعة المستخدمة من الملعب (فقط في جدول المباريات).
- إنشاء سمة host_country_code.
- إنشاء سمة رقمية tournament_year.
- إنشاء سمة full_name للاعب بتنسيق يمكن قراءته.
- إنشاء سمة short_stage_name والتي تتضمن دور خروج المغلوب ودور المجموعات فقط.
- إنشاء سمة winner_code في جدول البطولات.
- قم بإنشاء سمة ثنائية late_goal والتي تشير إلى ما إذا كان الهدف قد تم تسجيله في وقت متأخر من المباراة. اعتمد على دقيقة دخول الهدف وكذلك شوط المباراة مستفيدًا من خبرتك في المجال لتحديد الهوامش.
- قم بإنشاء أي سمة من الممكن أن تساعدك في عملية الاستكشاف أو المكاملة.

3. الاستكشاف والتحليل

a. دراسة حالة الحضور

- قم برسم المتوسط والوسيط باستخدام line chart وذلك ضمن نفس المخطط.
- قم بعرض توزع السمة على histogram مستخدمًا عددًا تراه مناسبًا للـ bins.
- قم بعرض توزع السمة ضمن كل دورة للبطولة باستخدام boxplot.

ماذا تستنتج من كل مخطط؟

b. دراسة حالة الأهداف

- ارسم متوسط (أو وسيط) فترة الهدف لكل نسخة للبطولة باستخدام bar chart
- قم بعرض histogram لإجمالي عدد الأهداف في المبارة الواحدة في كأس العالم.
- احسب دقيقة الهدف والمدة الزمنية الأكثر تكرارًا لكل نسخة للبطولة
- قم بعرض histogram لمجموع الأهداف المتأخرة في كل نسخة للبطولة
- قم باستخدام bar chart لعرض أفضل 12 هدافًا على الإطلاق في كأس العالم
- قم باستخدام bar chart لعرض أفضل هداف في كل نسخة للبطولة.
- قم باستخدام bar chart لعرض إجمالي عدد الأهداف في كل نسخة للبطولة.
- ضع في اعتبارك منتخبات البرازيل وألمانيا وإيطاليا، واستخدم strip plot لدقيقة الهدف واسم الدور (أو اسم الدور القصير).

ماذا تستنتج من كل مخطط؟

c. دراسة حالة المباريات

- احسب تكرارات المباريات عبر تاريخ بطولات كأس العالم، آخذًا بعين الاعتبار أن سمات home_team وaway_teamg هما سمات تبادلية في كأس العالم!
- ارسم أكثر 10 مباريات تكررت في كأس العالم باستخدام bar chart.

d. دراسة حالة البطولة

- عين مجموعة اللاعبين الذين مثلوا أكثر من فريق ثم حاول اكتشاف الأسباب الكامنة وراء هذه الظاهرة.
- هل هناك علاقة بين الدولة المضيفة والفائز بالبطولة؟
- هل هناك ارتباط بين مباراة المضيف وفئة نسبة الحضور الجماهيري؟
- هل هناك علاقة ارتباط بين الدولة المضيفة وفئة الحضور الجماهيري؟
- حاول شرح النتائج بالاستفادة من معرفتك بالمجال.

توجيه: يمكنك استخدام Cramer's V لقياس قوة الارتباط إن وجدت.

(انظر المصادر لمزيد من المعلومات)

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، واعادة تحقيق أي متطلب أو إجرائية ولكن عليك أن توضح السبب وراء إعادة التحقيق ضمن المفكرة. تذكر أن التنقيب هو عملية ديناميكية وتكرارية، وعليه، تأكد من كتابة كود نظيف وقابل للقراءة وإعادة الاستخدام وحاول تنظيم المفكرة قدر الإمكان.

تذكر 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

1. سيتم تسليم هذه الوظيفة على مرحلتين، حيث تغطي المرحلة الأولى أول متطلبين وتغطي المرحلة الثانية المتطلبات المتبقية.
2. يجب إكمال المتطلبين 1 و 2 في غضون أربعة أيام من إعلان الوظيفة.
3. يجب إكمال باقي الطلبات في غضون ثمانية أيام من إعلان الوظيفة.
4. سيتم الإعلان عن نموذج التقديم في موعد لاحق.
5. 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 ويحصل الطالب على درجة الصفر كاملةً!
6. 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.
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 أيضًا الثقة بين الطالب والمدرس ويمكن أن يؤدي إلى عواقب وخيمة، لذلك تجنب مشاركة عملك مع الآخرين أو السماح للآخرين باستخدام عملك.
8. لا تستخدم أي مادة أو مكتبة غير مصرح بها ما لم يسمح المدرس صراحة بذلك (يمنع استخدام matplotlib, seaborn على سبيل المثال لا الحصر!).
9. افهم الفرق بين التعاون والغش، يُسمح بالتعاون طالما أنه لا يتضمن جزءًا من عمل زملائك.
10. كن صريحًا وشفافًا بشأن عملك.
11. إذا كانت لديك أسئلة بشأن الوظيفة فاطلب التوجيه من المدرس.
12. قبل البدء بتحقيق الوظيفة اقرأ كل طلب بتمعن وكذلك انظر إلى كتيب قاعدة بيانات Fjelstul لتحصل على توصيف دقيق لكل عمود ضمن الجداول.
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 كاملةً!
16. لا يجب عليك التكلف في استنتاجاتك، اكتبها بأي طريقة تناسبك وبأي لغة تريدها (العربية أو الإنجليزية)، موضحًا 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 مناسبًا.

المصادر:

1. Fjelstul world cup database
2. World cup attendance dataset
3. World Cup Referees dataset
4. Data exploration notebook
5. World cup wiki
6. World cup hosts wiki
7. How to easily check for associations between categoricals