Data Science SS24 – Aufgabe 2

In diesem Projekt wirst du deine Fähigkeiten im Daten-Wrangling und in der explorativen Datenanalyse auf Baseball-Daten anwenden. Insbesondere möchten wir wissen, wie gut **Moneyball** für die Oakland A's funktioniert hat.

Eine kurze Erklärung wie Baseball gespielt wird gibt es hier:

https://www.youtube.com/watch?v=RlVm1qNhzXE&pp=ygUTYmFzZWJhbGwgZXJrbMOkcnVuZw%3D%3D

Ein bisschen Hintergrund

Wir werden uns Daten über Teams in der Major League Baseball ansehen. Ein paar wichtige Punkte:

Die Major League Baseball ist eine professionelle Baseballliga, in der Teams Spieler bezahlen, um Baseball zu spielen. Das Ziel jedes Teams ist es, so viele Spiele wie möglich aus einer 162-Spiele-Saison zu gewinnen. Teams gewinnen Spiele, indem sie mehr Läufe als ihr Gegner erzielen. Im Prinzip sind bessere Spieler teurer, sodass Teams, die gute Spieler wollen, mehr Geld ausgeben müssen. Teams, die am meisten ausgeben, gewinnen häufig auch am meisten. Also lautet die Frage, wie kann ein Team, das nicht so viel ausgeben kann, gewinnen? Die grundlegende Idee, die Oakland (und andere Teams) verwendet haben, ist, neu zu definieren, was einen Spieler gut macht, d.h. herauszufinden, welche Spielermerkmale in Siege umgesetzt werden. Sobald sie erkannten, dass Teams Spieler nicht wirklich anhand dieser Merkmale bewerteten, konnten sie dies ausnutzen, um unterbewertete Spieler zu bezahlen, Spieler, die nach ihren Maßstäben gut waren, aber von anderen Teams nicht als solche erkannt und daher nicht so teuer waren.

Du kannst mehr Informationen über diese Periode in der Baseballgeschichte von folgenden Quellen erhalten:

Wikipedia

Der Moneyball-Film

Die Daten

Du wirst Daten aus einer sehr nützlichen Datenbank über Baseball-Teams, Spieler und Saisons verwenden, die unter http://www.seanlahman.com kuratiert wurde. Die Datenbank wurde als **sqlite**-Datenbank unter https://github.com/jknecht/baseball-archive-sqlite zur Verfügung gestellt. **sqlite** ist ein leichtgewichtiges, dateibasiertes Datenbankmanagementsystem, das sich gut für kleine Projekte und Prototypen eignet.

Du kannst hier mehr über den Datensatz erfahren:

https://github.com/fonnesbeck/baseball/blob/master/data/lahman-csv_2015-01-24/readme2014.txt

Du kannst die **sqlite**-Datei direkt von GitHub herunterladen:
https://github.com/jknecht/baseball-archive-sqlite/releases/download/2022/lahman_1871-2022.sqlite

Du wirst auf die **sqlite**-Datenbank in Python mit dem sqlite-Paket zugreifen. Dieses Paket bietet eine unkomplizierte Schnittstelle, um Daten aus **sqlite**-Datenbanken mit standardmäßigen SQL-Befehlen zu extrahieren.

Sobald du eine Verbindung mit der **sqlite**-Datenbank hergestellt hast, kannst du Abfrageergebnisse direkt in einem pandas-DataFrame speichern, indem du die read_sql-Funktion verwendest.

Zum Beispiel, so würdest du die gesamten Ligagehälter für jedes Jahr tabellieren:

```
import sqlite3
import pandas

sqlite_file = 'lahman2014.sqlite'
conn = sqlite3.connect(sqlite_file)

salary_query = "SELECT yearID, sum(salary) as total_payroll FROM Salaries
WHERE lgID == 'AL' GROUP BY yearID"

team_salaries = pandas.read_sql(salary_query, conn)
team_salaries.head()
```

Das Ergebnis würde etwa so aussehen:

	yearID	total_payroll
0	1985	134401120.0
1	1986	157716444.0
2	1987	136088747.0
3	1988	157049812.0
4	1989	188771688.0

Die Frage

Wir möchten verstehen, wie effizient Teams historisch Geld ausgegeben und dafür Siege erzielt haben. Im Falle von Moneyball würde man erwarten, dass Oakland vor 2000 nicht viel effizienter in ihren Ausgaben war als andere Teams, zwischen 2000 und 2005 viel effizienter war (immerhin wurde darüber ein Film gemacht) und bis dahin andere Teams aufgeholt haben könnten. Deine Aufgabe in diesem Projekt ist es, zu sehen, wie dies in den Daten, die wir haben, reflektiert wird.

Teil 1: Wrangling

Die Daten, die du benötigst, um diese Fragen zu beantworten, befinden sich in den Tabellen Salaries und Teams der Datenbank.

Problem 1

Verwende SQL, um eine Beziehung zu berechnen, die das Gesamtgehalt und den Gewinnprozentsatz (Anzahl der Siege / Anzahl der Spiele * 100) für jedes Team (d.h. für jede TeamID- und YearID-Kombination) enthält. Du solltest andere Spalten einbeziehen, die bei der späteren Durchführung der Analyse hilfreich sein werden (z.B. Franchise-IDs, Anzahl der Siege, Anzahl der Spiele).

Füge den SQL-Code, den du verwendet hast, um diese Beziehung zu erstellen, in deinen Bericht ein. Beschreibe, wie du mit fehlenden Daten in diesen beiden Beziehungen umgegangen bist. Gib insbesondere an, ob in einer der Tabellen Daten fehlen und wie die Art des von dir verwendeten Joins bestimmt, wie du mit diesen fehlenden Daten umgegangen bist. Eine Anmerkung zu SQL: Du musst auf die Division von Ganzzahlen vs. Fließkommazahlen achten.

Teil 2: Explorative Datenanalyse

Gehaltsverteilung

Problem 2

Schreibe Code, um Diagramme zu erstellen, die die Verteilung der Gehälter über die Teams in Abhängigkeit von der Zeit (von 1990-2014) veranschaulichen.

Frage 1

Welche Aussagen kannst du über die Verteilung der Gehälter in Abhängigkeit von der Zeit aufgrund dieser Diagramme machen? Denke daran, dass du Aussagen in Bezug auf die zentrale Tendenz, die Streuung usw. machen kannst.

Problem 3

Schreibe Code, um Diagramme zu erstellen, die speziell mindestens eine der Aussagen, die du in Frage 1 gemacht hast, zeigen. Wenn du zum Beispiel eine Aussage darüber machst, dass es einen Trend gibt, dass die Gehälter im Laufe der Zeit abnehmen, erstelle ein Diagramm einer Statistik für die zentrale Tendenz (z.B. durchschnittliches Gehalt) vs. Zeit, um das speziell zu zeigen.

Korrelation zwischen Gehalt und Gewinnprozentsatz

Problem 4

Schreibe Code, um den gesamten Zeitraum in fünf Zeitperioden zu diskretisieren (du kannst pandas.cut verwenden, um dies zu erreichen) und dann ein Streudiagramm zu erstellen, das den durchschnittlichen Gewinnprozentsatz (y-Achse) vs. das durchschnittliche Gehalt (x-Achse) für jede der fünf Zeitperioden zeigt. Du könntest eine Regressionslinie (mit z.B. NumPys polyfit) in jedem Streudiagramm hinzufügen, um die Interpretation zu erleichtern.

Frage 2

Was kannst du über die Teamgehälter in diesen Perioden sagen? Gibt es Teams, die besonders gut darin sind, für Siege zu bezahlen, über diese Zeitperioden hinweg? Was kannst du über die Ausgabeneffizienz der Oakland A's in diesen Zeitperioden sagen (das Beschriften von Punkten im Streudiagramm kann bei der Interpretation helfen)?

Teil 3: Datentransformationen

Standardisierung über die Jahre hinweg

Es scheint problematisch zu sein, Gehälter über die Jahre hinweg zu vergleichen, also lass uns eine Transformation durchführen, die bei diesen Vergleichen hilft.

Problem 5

Erstelle eine neue Variable in deinem Datensatz, die das Gehalt in Abhängigkeit vom Jahr standardisiert. Diese Spalte für Team **i** im Jahr **j** sollte folgendermaßen gleich sein:

$$standardisiertes\_gehalt\_ij = (gehalt\_ij - durchschnittliches\_gehalt\_j) / s\_j$$

Wobei:

- **standardisiertes_gehalt_ij** die standardisierte Gehaltsspalte für Team i im Jahr j ist,
- **gehalt_ij** das Gehalt für Team i im Jahr j,
- **durchschnittliches_gehalt_j** das durchschnittliche Gehalt für das Jahr j, und
- **s_j** die Standardabweichung des Gehalts für das Jahr j ist.

Problem 6

Wiederhole die gleichen Diagramme wie in Problem 4, verwende aber diese neue standardisierte Gehaltsvariable.

Frage 3

Diskutiere, wie die Diagramme aus Problem 4 und Problem 6 die Transformation widerspiegeln, die du an der Gehaltsvariablen vorgenommen hast.

Erwartete Siege

Es ist schwer, globale Trends über Zeitperioden hinweg mit diesen mehreren Diagrammen zu sehen, aber jetzt, da wir die Gehälter über die Zeit standardisiert haben, können wir uns ein einzelnes Diagramm ansehen, das die Korrelation zwischen Gewinnprozentsatz und Gehalt über die Zeit zeigt.

Problem 7

Erstelle ein einzelnes Streudiagramm des Gewinnprozentsatzes (y-Achse) vs. standardisiertes Gehalt (x-Achse). Füge eine Regressionslinie hinzu, um die Beziehung hervorzuheben.

Die Regressionslinie gibt dir den erwarteten Gewinnprozentsatz als Funktion des standardisierten Gehalts. Aus der Betrachtung der Regressionslinie scheint es, dass Teams, die ungefähr das durchschnittliche Gehalt in einem gegebenen Jahr ausgeben, 50% ihrer Spiele gewinnen werden (d.h. der Gewinnprozentsatz ist 50, wenn das standardisierte Gehalt 0 ist), und Teams erhöhen 5% Siege für jede 2 Standard-Einheiten des Gehalts (d.h., der Gewinnprozentsatz ist 55, wenn das standardisierte Gehalt 2 ist). Wir werden sehen, wie dies im Allgemeinen mit linearer Regression später im Kurs gemacht wird.

Aus diesen Beobachtungen können wir den erwarteten Gewinnprozentsatz für Team **i** im Jahr **j** als

$$erwartete\_siegquote\_ij = 50 + 2{,}5 \times standardisiertes\_gehalt\_ij$$

Ausgabeneffizienz

Mit diesem Ergebnis können wir nun ein einzelnes Diagramm erstellen, das es einfacher macht, die Effizienz der Teams zu vergleichen. Die Idee ist, eine neue Maßeinheit für jedes Team zu erstellen, die auf ihrem Gewinnprozentsatz und ihrem erwarteten Gewinnprozentsatz basiert, die wir über die Zeit hinweg in einem Diagramm darstellen können, das zusammenfasst, wie effizient jedes Team in seinen Ausgaben ist.

Problem 8

Erstelle ein neues Feld, um die Ausgabeneffizienz jedes Teams zu berechnen, gegeben durch

$$effizienz\_ij = siegquote\_ij - erwartete\_siegquote\_ij$$

für Team **i** im Jahr **j**.

Erstelle ein Liniendiagramm mit dem Jahr auf der x-Achse und der Effizienz auf der y-Achse. Eine gute Auswahl an Teams zum Plotten sind Oakland, die New York Yankees, Boston, Atlanta und Tampa Bay (Team-IDs OAK, BOS, NYA, ATL, TBA).

Frage 4

Was kannst du aus diesem Diagramm im Vergleich zu den Diagrammen lernen, die du in Frage 2 und 3 betrachtet hast? Wie gut war Oaklands Effizienz während der Moneyball-Periode?

Einreichung

Bereite ein Jupyter-Notebook das für jedes Problem Folgendes enthält:

(a) Code, um den besprochenen Schritt durchzuführen,

(b) Ausgabe, die das Ergebnis deines Codes zeigt, und

(c) eine kurze Prosabeschreibung, wie dein Code funktioniert.

Denke daran, die von dir vorbereitete Ausarbeitung soll deine Datenanalyse effektiv kommunizieren. Gedankenlos große Mengen von Ausgaben in deiner Ausarbeitung zu zeigen, widerspricht diesem Zweck.

Alle Achsen in Diagrammen sollten auf informative Weise beschriftet sein. Deine Antworten auf jede Frage, die sich auf ein Diagramm bezieht, sollten sowohl

(a) eine Textbeschreibung deines Diagramms als auch

(b) ein oder zwei Sätze Interpretation in Bezug auf die gestellte Frage enthalten.