Hochschule für Wirtschaft und Recht Berlin
Fachbereich 1 Wirtschaftswissenschaften

WiSe 2023-2024

Dozent: **Dr. Alessandro Bramucci**
Seminar: **Einführung in Python für Data Analytics**

Abschlussproket

Zielsetzung

Sie sind der neu eingestellte Datenanalyst bei dem vor kurzem gegründeten Unternehmen Elektratuto AG. Das Unternehmen vertreibt bundesweit Elektroautos und ist in fast allen Bundesländern vertreten. Das Unternehmen ist erst seit ein paar Monaten im Geschäft, hat aber bereits eine Vielzahl von Autos verkauft. Der Vorstand des Unternehmens ist daran interessiert, mehr zu erfahren, warum Kunden ihre Autos kaufen, die sozioökonomischen Merkmale seiner Kundschaft sowie die Faktoren, die zum Kauf des Autos geführt haben. Zu diesem Zweck hat das Unternehmen eine Umfrage unter seinen Kunden durchgeführt. Die Umfrage wurde unter allen Personen durchgeführt, die das Auto ausprobiert haben. Allerdings hat sich nicht jeder, der das Auto getestet hat, auch zum Kauf entschlossen.

Präsentation

Ihre Aufgabe ist es daher, die gesammelten Daten zu analysieren und einige Fragen zu beantworten. Gemeinsam mit Ihrem Vorgesetzten haben Sie beschlossen, eine Liste von 20 Problemen bzw. Fragen zu bearbeiten. Die Liste finden Sie auf der nächsten Seite. Es ist wichtig, dass Sie alle 20 Aufgaben bzw. Fragen beantworten. Sie können entscheiden, in welcher Form die Informationen dargestellt werden sollen, z.B. mit einer Abbildung (Kuchendiagramme, Balkendiagramme, Streudiagramme, usw.), einer Tabelle oder Stichpunkten. Aus der Präsentation (max. 20 Minuten) muss klar erkennbar sein, auf welche Frage oder Aufgabe die Informationen sich beziehen. Es gibt nur eine einzige Bedingung. Sie müssen mindestens 5 Abbildungen präsentieren. Auch der CEO von Elektratuto AG wird bei Ihrer Präsentation anwesend sein, zusammen mit Ihren Kollegen aus der Datenabteilung und Ihrem Vorgesetzten. Denken Sie daran, dass der CEO sehr beschäftigt ist und nur 5 Minuten Zeit hat, sich Ihre Präsentation anzuschauen. Der Rest der Präsentation wird vor Ihrem Vorgesetzten sowie Ihre Kollegen gehalten. Es liegt an Ihnen, zu entscheiden, welche Informationen Sie dem CEO und den Datenexperten präsentieren wollen. Denken Sie daran, dass der CEO weder ein Python noch ein Datenexperte ist.

Daten

Die Daten wurden in zwei separaten Dateien gesammelt. Die erste Datei (*kunden.csv*) enthält die Ergebnisse der Befragung unter den Personen, die das Auto gekauft haben. In diesem Datensatz sind dann Informationen über den Kaufpreis des Fahrzeugs verfügbar. Wundern Sie sich nicht, wenn die Preise stark variieren. Die Elektrauto AG verkauft viele verschiedene Modelle, die auch stark individuell angepasst werden können. Die zweite Datei (*besucher.csv*) enthält Informationen von Personen, die das Auto ausprobiert haben aber nicht gekauft. Vorsicht! Die Umfrage war komplett anonym. Es ist jedoch möglich, dass einige Kunden oder Besucher keine persönlichen Informationen angeben wollten. Es ist daher möglich, dass einige Informationen falsch sind. Es gibt ein weiteres Problem. Die Informationen über das geografische Gebiet (Bundesland) des Händlers wurden in einer separaten Datei erfasst (*geo.txt*). Außerdem scheint es eine Reihe von Unstimmigkeiten zu geben, wie die Daten gespeichert wurden, z.B. wurde manchmal 'Nordrhein-Westfalen' als 'NRW' gespeichert. Ihre Aufgabe ist es, die Datensätze zu bereinigen und die Dateien zusammenzuführen.

Fragen und Aufgaben

1. Wie viele Autos wurden verkauft?
2. Was ist der Höchst-, Mindest- und Durchschnittspreis der verkauften Autos?
3. Wie hoch war der Gesamtumsatz?
4. Wie viele Autos wurden pro Bundesland verkauft?
5. Wie hoch war der durchschnittliche Umsatz pro Bundesland?
6. Haben mehr Frauen oder mehr Männer unsere Autos gekauft?
7. Wie hoch ist das Durchschnittsalter unserer Kunden?
8. Wie hoch ist das Durchschnittsalter der Besucher in unseren Showrooms?
9. Was ist das Durchschnittsalter unserer männlichen Kunden?
10. Wie hoch ist das Durchschnittseinkommen unserer Kunden?
11. Wie hoch ist die Korrelation (Pearson-Korrelation) zwischen den Variablen Alter, Einkommen, Preis und Zeit? (nur für die Kunden)
12. Wie ist die Variable Zeit verteilt? (Kunden und Besucher zusammen)
13. Wie viele Kunden haben keinen Kredit bei der Bank genommen, um das Auto zu kaufen? Die Kundenabteilung hat vergessen, diese Informationen zu sammeln. Wir können davon ausgehen, dass die Kunden mit einem Jahreseinkommen, das höher als der Autopreis ist, keinen Bankkredit benötigten.

14. Welches sind die sozioökonomischen Merkmale der Kunden, die den Kaufpreis beeinflussen? Wählen Sie die geeigneten abhängigen Variablen aus und schätzen Sie eine Regression unter Verwendung der geeigneten Methode.

15. Prognostizieren Sie den Kaufpreis eines unserer Autos für einen männlichen Kunden im Alter von 32 Jahren mit einem Einkommen von 30.000 Euro. Prognostizieren Sie den Kaufpreis eines unserer Autos für einen männlichen Kunden im Alter von 51 Jahren und mit einem Einkommen von 54.000 Euro.

16. In Bezug auf die vorherige Frage: Welche Variable beeinflusst den Preis des Autos am meisten? Mit anderen Worten: Die von Ihnen geschätzten Regressionskoeffizienten müssen direkt vergleichbar sein. Wie sollen die Daten transformiert werden? Tipp: Beta-Werte.

17. Schätzen Sie eine Regression, die die Wahrscheinlichkeit des Kaufs eines Autos ermittelt. Verwenden Sie die entsprechende Methode. Tipp: Logistische Regression.

18. Wie hoch ist die Wahrscheinlichkeit, dass ein 32-jähriger männlicher Kunde mit einem Einkommen von 30.000 Euro, der das Auto 30 Minuten lang getestet hat, eines unserer Modelle kauft? Wie hoch ist die Wahrscheinlichkeit, dass ein 51-jähriger männlicher Kunde mit einem Einkommen von 54.000 Euro, der das Auto 45 Minuten lang getestet hat, eines unserer Modelle kauft?

19. Auf welche Probleme sind Sie bei der Zusammenführung des Datensatzes gestoßen? Stellen Sie die Operationen vor, die Sie zum Zusammenführen und Bereinigen der Daten durchgeführt haben.

20. Welche Vorschläge würden Sie der Kundenabteilung für die Umfrage im nächsten Jahr machen? Welche zusätzlichen Informationen sollten gesammelt werden? Formulieren Sie zwei Vorschläge.

Datensätze und Variablenbeschreibung

1) *kunden.csv*

Name	Beschreibung
Alter	Alter des Kunden. Jahren
Einkommen	Jahreseinkommen des Kunden. Euro
Preis	Kaufpreis. Euro
Geschlecht	1 für männlich; 0 für weiblich
Zeit	Fahrzeug-Testzeit. Minuten
KundeNr	Kundennummer

2) *besucher.csv*

Name	Beschreibung
Alter	Alter des Kunden. Jahren
Einkommen	Jahreseinkommen des Kunden. Euro
Zeit	Fahrzeug-Testzeit. Minuten
Geschlecht	1 für männlich; 0 für weiblich
KundeNr	Kundennummer

3) *geo.txt*

Name	Beschreibung
KundeNr	Kundennummer
Niederlassung	Bundesland der Händler

Abgabe und Bewertung

Alle Gruppen müssen das Projekt bis **Donnerstag, den 4. Januar um 23:59 Uhr** abschließen. Ein Vertreter der Gruppe sollte mir bis zum oben genannten Datum eine E-Mail mit dem Link zum GitHub-Repository zusammen mit dem Link zur Online-Präsentation schicken.

Das Projekt besteht aus 20 Aufgaben. Für jede richtige Antwort erhält die Gruppe einen Punkt. Die Projektnote wird mit einem Gewicht von 50 Prozent in die Endnote einfließen. Weitere drei Punkte werden für die Qualität der Präsentation, den Stil des Codes und die Zusammenarbeit auf GitHub vergeben. Diese drei Punkte werden in der Endnote mit 10 Prozent gewichtet. Die verbleibenden 40 Prozent sind für die mündliche Prüfung vorgesehen (10 Prozentpunkte für jede richtige Antwort). Wir machen ein Beispiel, um die Berechnung der Endnote zu erklären. Nehmen wir an, dass die Gruppe alle 20 Punkte des Projekt (20/20) sowie die 3 Punkte für Codes, GitHub und Präsentation (3/3) erhalten hat. Die Gruppe hat dann 60 Prozentpunkte erreicht.

$$\frac{20}{20} \cdot 0,5 + \frac{3}{3} \cdot 0,1 = 0,6$$

Die drei Punkte für Code usw. werden auf der Basis folgender Kriterien vergeben:
- Die Folien sind gut gegliedert und die Informationen werden klar präsentiert.
- Die Präsentation bleibt innerhalb des vorgegebenen Zeitrahmens.
- Die Arbeit wurde gleichmäßig zwischen den Gruppenteilnehmern verteilt.
- Der Code ist gut organisiert und umfangreich kommentiert.
- Der Code kann auch auf einem anderen Computer ohne Fehler ausgeführt werden.
- Git und GitHub wurden für die Zusammenarbeit an dem Projekt verwendet, z.B. *branch* and *pull requests*.

Berlin, 15.12.2023