Hochschule für Wirtschaft und Recht Berlin
Fachbereich 1 Wirtschaftswissenschaften

SoSe 2024

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

Abschlussproket

Zielsetzung

Sie sind der neu Datenanalyst bei Elektratuto AG, ein vor kurzem gegründeten Autohersteller im Bereich der E-Autos. Das Unternehmen vertreibt bundesweit seine Automodelle und ist in fast allen Bundesländern vertreten. Das Unternehmen ist erst seit ein paar Monaten im Geschäft, hat aber bereits eine ziemlich große Menge an Autos verkauft. Der Vorstand hat großes Interesse zu erfahren, warum die Kunden ihre Autos kaufen. Insbesondere möchte der Vorstand die sozioökonomischen Merkmale seiner Kundschaft untersuchen sowie die Faktoren, die zum Kauf eines Autos geführt haben. Zu diesem Zweck hat das Unternehmen eine Umfrage unter seinen Kunden durchgeführt. Die Umfrage wurde unter allen Personen verteilt, die das Auto ausprobiert haben. Allerdings hat sich nicht jede Person, die das Auto getestet hat, auch zum Kauf entschlossen.

Präsentation

Ihre Aufgabe ist es, die Ergebnisse der Umfrage auszuwerten. 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 Fragen beantworten. Sie können selbst entscheiden, in welcher Form die Informationen dargestellt werden sollen, z.B. mit einer Abbildung (Kuchendiagramme, Balkendiagramme, Streudiagramme, usw.), mit einer Tabelle oder mit einem Text (oder Stichpunkten). Aus der Präsentation (max. 20 Minuten) muss aber klar erkennbar sein, auf welche Frage Sie sich beziehen. Markieren Sie dann bitte jede Antwort mit einer kurzen Bezeichnung, z.B. F1 für die erste Frage, F2 für die zweite Frage usw. 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 welche Informationen Sie 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 Kunden, also die Personen die das Auto getestet und gekauft haben. In diesem Datensatz sind Informationen über den Kaufpreis des Fahrzeugs verfügbar. Wundern Sie sich nicht, wenn die Preise sehr 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 aber nicht gekauft haben. 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 oder fehlerhaft sind. Das bedeutet aber nicht dass das Auto nicht getestet oder verkauft wurde! Es gibt ein weiteres Problem. Die Informationen über das Bundesland des Händlers wurden in einer separaten Datei erfasst (*geo.txt*). Hier leider wurden die Informationen in unterschiedlichen Formaten gespeichert, z.B. das Bundesland Nordrhein-Westfalen wurde manchmal als 'Nordrhein-Westfalen' und manchmal als 'NRW' eingetragen. Ihre Aufgabe ist es, die Datensätze zu bereinigen und zusammenzuführen.

Fragen

1. Wie viele Autos wurden verkauft?[1]
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. In welchem Bundesland wurde das teuerste Auto verkauft?
7. Haben mehr Frauen oder mehr Männer unsere Autos gekauft?
8. Wie hoch ist das Durchschnittsalter unserer Kunden? Wie hoch ist das Durchschnittsalter unserer Besucher?
9. Wie hoch ist das Durchschnittseinkommen unserer Kunden? Wie hoch ist das Durchschnittseinkommen unserer Besucher?
10. Bestimmen Sie, ob es einen statistisch signifikanten Unterschied zwischen dem Durchschnittseinkommen der Kunden und dem Durchschnittseinkommen der Besucher gibt.[2]

[1] Tipp: 1104.

[2] Sie müssen einen Zweistichproben-t-Test durchführen. Beginnen Sie mit der Aufstellung der Null- und Alternativhypothese. Verwenden Sie ein Signifikanzniveau (α) von 5%. Nehmen Sie an, dass die Varianzen der beiden Gruppen gleich sind.

11. Wie hoch ist die Korrelation (Pearson-Korrelation) zwischen den Variablen Alter, Einkommen, Preis und Zeit?[3]

12. Testen Sie grafisch mit Hilfe eines qq-Plots, ob die Variable Zeit approximativ normalverteilt ist.[4]

13. Wie viele Kunden haben keinen Bankkredit aufgenommen, um das Auto zu kaufen? Die Kundenabteilung hat vergessen, diese Informationen zu erfassen. Wir können davon ausgehen, dass die Kunden mit einem Jahreseinkommen höher als der Autopreis keinen Kredit aufgenommen haben.

14. Welche sozioökonomischen Merkmale beeinflussen den Kaufpreis? Wählen Sie die geeigneten unabhängigen Variablen aus und schätzen Sie eine Regression unter Verwendung der geeigneten Methode.[5]

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?[6]

17. Schätzen Sie eine Regression, die die Wahrscheinlichkeit des Kaufs eines Autos ermittelt. Verwenden Sie die entsprechende Methode.[7]

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.

[3] Berechnen Sie die Korrelation nur für Kunden.
[4] Kunden und Besucher zusammen.
[5] Verwenden Sie die Bibliothek *statsmodel* und nicht *sklearn*.
[6] Sie müssen standardisierten Koeffizienten schätzen.
[7] Sie müssen eine logistische Regression durchführen.

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 27. Juni um 23:59 Uhr** abgeschloßen haben. Die Präsentation muss mit Jupyter-Notebook erstellt werden. Eine Vertreterin oder Vertreter der Gruppe sollte mir bis zum oben genannten Datum eine E-Mail mit dem Link zum GitHub-Repository und dem Link zur Online-Präsentation schicken. Bitte setzen Sie die Teilnehmer Ihrer Gruppe in *cc*. Sobald die Korrekturen abgeschlossen sind, werde ich auf diese E-Mail die Bewertung schicken.

Das Projekt besteht aus 20 Aufgaben. Für jede richtige Antwort erhält die Gruppe einen Punkt. Die Punkten im Projekt werden dann 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).

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

Die drei Punkte für Code usw. werden auf der Basis folgender Kriterien vergeben:
- Die Folien sind gut gegliedert und die Aufgaben klar präsentiert.
- Die Antworten sind mit der Beschriftung F1, F2, usw. gekennzeichnet.
- Die Präsentation bleibt innerhalb der vorgegebenen Zeit.
- Die Präsentation wurde online veröffentlicht.
- Die Arbeit wurde gleichmäßig zwischen den Gruppenteilnehmern verteilt.
- Der Code ist gut organisiert und umfangreich kommentiert.
- Der Code kann auch auf einem anderen Computer fehlerfrei ausgeführt werden.
- Git und GitHub wurden für die Zusammenarbeit genutzt (z.B. *branches*, *pull requests*).

Berlin, 06.06.2024