Praxistage zur statistischen Datenauswertung

Aufgabeneinführung:

Die Universität „Michael Scott's University of Excellence" hat ihre Studenten an allen deutschen Standorten zur Teilnahme an der alljährlichen Studentenumfrage gebeten. Als neues Mitglied des DataAnalyticsTeam® der Universität besteht nun deine Aufgabe darin, diesen Datensatz auszuwerten. Das Hauptziel dieser Auswertung besteht darin, die vorgegebene abhängige Variable bestmöglich aufzuklären (= das bestmögliche R^2 zu erreichen, welches möglich ist). Deine abhängigen Variablen sind dabei folgende vier: Lebenszufriedenheit, Studienzufriedenheit, Mathenote im 2. Semester, Zufriedenheit mit der Note.

Datensatzbeschreibung:

Im Datensatz befinden sich 32 Variablen, welche für die Analyse genutzt werden sollen. Hierbei ist die abhängige Variable je nach zufälliger Gruppenzugehörigkeit anders, während die restlichen 31 Variablen im Datensatz für alle Gruppen vollständig identisch sind. Leere Zellen im Datensatz entsprechen fehlenden Werten. Alle Variablen wurden bereits vorcodiert (= alle Variablen zeigen nur noch Zahlenausprägungen), das mitgelieferte „Codebook" zeigt für jede Frage im Fragebogen an, welche Ausprägungen hinter den jeweiligen Zahlen im Datensatz stehen.

Vorgehen und Aufgabenstellung:

1. Führe für ALLE VARIABLEN im Datensatz einen einzelnen statistischen Test mit der ausgewählten abhängigen Variable durch und erstelle einen Datensatz, in welchem nur noch die Variablen enthalten sind, welche in den einzelnen Tests ein signifikantes Ergebnis gezeigt haben. (Hierdurch dürfte der Ausgangsdatensatz einige Variablen verlieren und kleiner werden) Als Tipp zur Dokumentation: Eine Tabelle mit den Spalten „Variable", „durchgeführter Test", „Ergebnis des Tests" ist ein guter Start!
2. Erstelle für diesen neuen Datensatz eine vollständige deskriptive Statistik und zeige dabei: Welche Variablen sind noch enthalten, wie sind diese ausgeprägt (Häufigkeiten / Lage- und Streuungsmaße / Boxplots & Histogramme)?
3. Zeige für 5 ausgewählte Variablen (diese sind frei von dir aus dem verkleinerten Datensatz mit den signifikanten Variablen zu wählen) eine vollständige Hypothesenauswertung mit allen in der Vorlesung festgehaltenen Schritten. Als Tipp für die Dokumentation: Ein vollständiger Output in Python kann hier wunderbar als Begleitmittel für die anstehende Präsentation dienen. Das Hypothesenpaar und das Fazit aus dem Test können die Folie abrunden. Aber präsentiert nicht einfach Eure Notebooks!
4. Erstelle aus allen Variablen ein gemeinsames Regressionsmodell für die Vorhersage der abhängigen Variable und interpretiere dies in den Grundlagen vollständig (R^2, adj. R^2, F-Statistik) sowie an 2-3 ausgewählten Variablen der Koeffizienten-Tabelle. Dieses Modell stellt nun dein neues Grundmodell dar, welches du gern verbessern möchtest. Tipp für die Präsentation: Screenshot!
5. Verbessere dieses Modell, sodass es die beste Statistik liefert. Achte hierbei darauf, dass du dein adj. R^2 maximieren willst. Das höchste adj. R^2 ist also perfekt für die

Vorhersage! Achte weiter darauf, wie Variablen in der Regression behandelt werden: Können also wirklich alle Variablen einfach eingesetzt werden oder muss man Python an manchen Stellen noch beibringen, dass sich innerhalb einer Variable Kategorien befinden zum Unterscheiden und nicht einfach Zahlen ausgewertet werden können? Beispiel: Geschlecht 1 und 2 versteht Python nicht als Kategorie; Geschlecht „Männlich“ und „Weiblich“ jedoch durchaus!

6. Stelle dieses bis zum Endpunkt verbesserte Modell zur Vorhersage deiner abhängigen Variable dar und interpretiere hierbei alle Bestandteile des Modells (sowohl Grundlagen, die vollständige Regressionsgleichung mit allen Koeffizienten sowie die Voraussetzungen des Modells). Zeige beispielhaft an dieser Regressionsgleichung, wie man nun einen Wert mit dieser Gleichung vorhersagen kann!
7. Fasse alle bisherigen Schritte in einer ~15-minütigen Präsentation zusammen die erklären soll, welche Gedanken hinter den einzelnen durchgeführten Schritten stehen, welche Interpretationen durchgeführt worden sind und wie das finale Ergebnis ist. Gehe in dieser Präsentation auf alle 6 vorher durchgeführten Punkte ein!

Allgemeine Hinweise für die Aufgabe:

Achte darauf, für deine Variable die jeweils passenden Tests auszuwählen -> Hierbei sind T-Tests, Korrelationen, Regressionen und auch die Varianzanalyse nötig. Recherchiere die Voraussetzungen, die Daten erfüllen müssen, um die bedenkenlose Anwendung von Regression und verschiedenen Hypothesentests zu erlauben. Prüfe, welche Voraussetzungen hier erfüllt oder nicht erfüllt sind, schätze die Risiken ein und schlage Verbesserungsmaßnahmen vor.

Die Auswahl der Tests bezieht sich hierbei auf das Skalenniveau; die abhängige Variable kann in allen Gruppen als mindestens intervallskaliert behandelt werden.

Da zum Schluss ALLE SCHRITTE der Aufgabenstellung innerhalb einer gemeinsamen Präsentation dargestellt werden soll, solltest du bereits von Anfang an die einzelnen Arbeitsschritte gut dokumentieren! Und: Ihr habt diesmal ein **klares Publikum**, dem ihr Eure Ergebnisse präsentiert. Es handelt sich um die Leitung der Hochschule, die sich ein Bild von ihrer Studentenschaft verschaffen will. Außerdem gibt es zwei Ziele der Universität, die lauten: 1) Ausbau der Exzellenz durch ausgezeichnete Studenten und 2) Schaffung einer „lebenswerten“ Universität durch eine höhere Zufriedenheit der Studentenschaft. Auch hier erhofft sich die Leitung insgeheim Tipps von Euch.

Letzte Hinweise:

Du bist bereits Profi und könntest diese Aufgabe in 1 Stunde erledigen? Super; lasse aber gern deiner Gruppe die Chance, selbst auf die Lösung zu stoßen und unterstütze diese nur, indem du passende Tipps gibst und nicht direkt die Endlösung vorgibst.