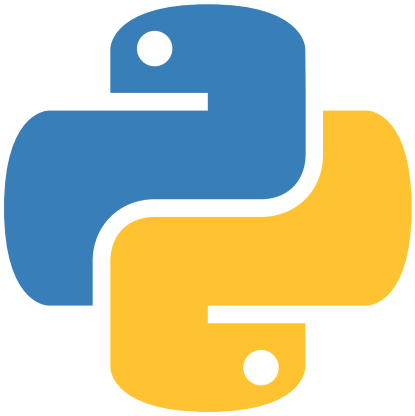
Business Case

Python



Beste kandidaat,

Gefeliciteerd met de voortgang naar de laatste ronde in het sollicitatieproces! Vóór je ligt de business case Python als onderdeel van het sollicitatieproces voor ons Data Science traineeship. We hebben je eerder gevraagd je te verdiepen in de wereld van Business Intelligence en haar meest belangrijke facetten. In deze business case word je gevraagd je opgedane kennis en vaardigheden in Python te gebruiken om een aantal vraagstukken op te lossen. Concreet zul je worden getest op je vermogen om data te analyseren en (basis) code te schrijven in Python. Dit doe je met een door ons aangeleverde dataset die informatie bevat over de streamingsdienst Netflix. Je schrijft je eigen code en beantwoordt vragen over de dataset. Denk ook na over waarom je bepaalde keuzes hebt gemaakt zodat je ons hier in de presentatie in mee kan nemen.

Veel succes!

Gebruikte dataset

- netflix_titles.csv

Inleiding

In deze business case ga je aan de slag met het schrijven van Python code in een Jupyter Notebook omgeving. Dit is een interactieve omgeving waarin je zowel met code als geschreven tekst aan de slag kunt. Je werkt met een door ons aangeleverde dataset die je gaat analyseren en visualiseren. De bedoeling is dat je een aantal vragen over de data gaat beantwoorden aan de hand van jouw gegenereerde code. Bij het beantwoorden van de vragen zijn we geïnteresseerd in jouw antwoord op de vraag, jouw code en eventueel een korte verklaring over hoe je tot je antwoord bent gekomen.

Stappenplan Business Case

Voor deze opdracht werken we met Anaconda Navigator en Jupyter Notebook. Anaconda Navigator is een grafische gebruikersomgeving waarin je toegang hebt tot verschillende programmeertalen zoals Python en R. Ook heb je toegang tot Jupyter Notebook: een interactieve web applicatie waarin live code en tekst samenkomen. Met het downloaden van Anaconda zul je automatisch toegang krijgen tot de Anaconda Navigator en Jupyter Notebook.

Stap 1

Ga naar https://www.anaconda.com/products/distribution en scroll naar beneden. Onder ‘Anaconda Installers’ kies je de 64-Bit Graphical Installer voor jouw besturingssysteem. Het downloaden van Anaconda vereist ongeveer 3.7 GB aan vrije ruimte. Doorloop de installatiewizard en laat verder alle instellingen ongewijzigd. Open Anaconda Navigator. Als je de melding krijgt dat er een nieuwere versie van Anaconda beschikbaar is, download je deze eerst. Je hebt nu succesvol Anaconda gedownload.

Stap 2

Open nu Anaconda Navigator. Je ziet een aantal tegels met functionaliteiten. Ga op zoek naar Jupyter Notebook en klik op ‘Launch’. Er wordt nu vanuit de browser een lokale versie van Jupyter Notebook gestart.

Stap 3

Maak een nieuwe map aan door rechts bovenin op ‘new’ te drukken en vervolgens op ‘folder’. Er wordt nu een nieuwe map aangemaakt genaamd ‘Untitled Folder’. Vink deze map aan en druk vervolgens links bovenin op ‘Rename’. Noem deze map ‘Business case DS’. Open de map en druk op ‘Upload’. Open ‘business_case_DS.ipynb’ en ‘netflix_titles.csv’ zoals je deze van ons hebt ontvangen. Je hebt nu toegang tot de dataset én het notebook bestand met de opdrachten.

Stap 4

Open het notebook bestand. Hier vind je begeleidende tekst, opdrachten en ruimte voor jouw code en/of antwoord. Doorloop het notebook bestand en sla het op zodra je alle vragen hebt beantwoord.

Hieronder kun je alvast bekijken welke kolommen de dataset bevat en wat deze kolommen precies inhouden. Ook vind je na de dataset tabel alvast een overzicht van alle opdrachten en vragen zoals deze ook in het notebook bestand staan. Deze hoef je in dit bestand niet te beantwoorden. Je kunt nu starten in 'business_case_DS.ipynb' en daar al je code en antwoorden noteren.

Dataset overzicht

Tabel 1. Netflix_titles dataset (12 kolommen en 8807 rijen)

Column name	Elaboration
showID	Primary key column
type	Movie or tv show
title	The title of the production
director	The director of the production
cast	The full cast of the production
country	The country in which the movie of tv show has been produced
date_added	The date the production was added to Netflix
release_year	The year the production was released
rating	The TV parental guidelines rating
duration	The duration of the production (in minutes and seasons)
listed_in	The genres the production is listed in
description	The productions' storyline description

Vragen

Exploratory Data Analysis (EDA) & Preprocessing

Q1: Schrijf code om de dataset (netflix_titles.csv) in te lezen. Creëer hiervoor een pandas dataframe genaamd 'netflix_df'. Bekijk de eerste 10 rijen van de dataframe.

Q2: Voer een initiële analyse uit op de dataset. Welke eerste algemene informatie haal je uit de dataset? Denk hierbij ook aan missende waarden, duplicaten en data types. Vervang alle missende waarden door lege strings.

Q3: In de 'rating' kolom staan 3 afwijkende waarden. Lokaliseer deze waarden en verwijder vervolgens de rijen die deze waarden bevatten. Laat de lege strings staan.

Q4: De waarden van een aantal kolommen staan nog in onhandige formats. Schrijf voor de volgende vragen code zodat de kolommen zich beter lenen voor visualisatie.

Q4.1: De kolommen 'title', 'cast', 'country', 'director' en 'description' worden niet gebruikt in deze opdracht. Verwijder deze.

Q4.2: De kolom 'date_added' staat in een onhandig format voor analyse (Month DD, YYYY). Maak twee nieuwe kolommen (year_added en month_added) die respectievelijk enkel het jaar en de maand bevatten van de kolom 'date_added'.

Q4.3: De kolom 'listed_in' bevat in veel gevallen meer dan 1 waarde. Dit is onhandig voor de visualisatie. Maak daarom een nieuwe kolom (first_genre) aan in de netflix_df die voor de originele kolom alleen de eerste waarde bevat. Doe dit door de strings te splitsen (.split methode).

Q4.4: De kolom 'rating' bevat informatie over leeftijdsgebonden content van Netflix. Deze kijkcijfer codes zijn echter niet heel betekenisvol. Maak een dictionary (rating_audience) aan met keys (ratings) en values (labels) o.b.v. de onderstaande codering. Zorg er tot slot voor dat 'rating_audience' een nieuwe kolom wordt in de netflix_df.

1.Everyone (‘TV-G’, ‘G’)
2.Younger kids (‘TV-Y’)
3.Older kids (‘TV-Y7’, ‘TV-Y7-FV’, ‘TV-PG’, ‘PG’)
4.Teens (‘TV-14’, ‘PG-13’)
5.Adults (‘TV-MA’, ‘R’, ‘NC-17’)
6.Unknown (‘UR’, ‘NR’)

Visualisatie

Q5: Wat is het aantal films versus tv shows?

> Q5.1: Schrijf code om het aantal films en tv series te tellen. Print het antwoord in de output.
>
> Q5.2: Gebruik matplotlib om een piechart te maken die het percentage films en tv series laat zien. Geef de plot de volgende titel:'Content type distribution' en een fontsize van 15.

Q6: In welk jaar zijn de meeste films uitgekomen? En in welk jaar de meeste tv shows? Maak m.b.v. Seaborn een plot die het aantal uitgekomen films en tv series per jaar (year_added) en per maand (month_added). Neem enkel de 10 meest populaire productiejaren mee.

Tip: Gebruik de parameter 'hue' om te differentiëren tussen films en tv series.

Q7: Welk genre komt het meest voor over alle producties van Netflix? Maak een variabele (genre_top20) met de 20 meest voorkomende genres en visualiseer deze m.b.v. een seaborn barplot. Gebruik hiervoor de eerder gemaakte kolom 'first_genre'.

Q8: Voor welk publiek is het grootste gedeelte van Netflix' aanbod geschikt? Gebruik je favoriete library om een simpele horizontale barplot te maken die de content distributie per doelgroep weergeeft. Gebruik hiervoor de eerder gemaakte kolom 'rating_audience'.