WPROWADZENIE DO NARZĘDZI DANYCH

Ludzie mają różne motywacje do realizowania tego, co ich interesuje. Zapytaj kogoś o samochód, a może powiedzieć, że nienawidzi sedanów, kocha SUV-y, albo nigdy nie dostałby niczego innego niż samochód elektryczny, a może w ogóle nie dostałby samochodu! Ludzie mają różne preferencje i nie zmienia się to w przypadku narzędzi do nauki o danych (statystycznych). Niektórzy ludzie kochają Excela do tego stopnia, że nie będą używać niczego poza tym oprogramowaniem do wszystkiego, od utrzymywania budżetu po analizę danych. Istnieje wiele powodów, dla których warto zachować poświęcenie, ale głównym powodem jest zapoznanie się z obiektem. Osoba, która prowadziła tylko drążek zmiany biegów, uwielbia sprzęgło, podczas gdy ci, którzy nigdy nie napędzali drążka, nie będą tak skłonni do preferowania ręcznej zmiany biegów. Jakie są powody preferowania jednej aplikacji od drugiej? Z mojego doświadczenia wynika, że istnieją trzy główne punkty:

1. Oprogramowanie jest łatwe w użyciu

2. Oprogramowanie jest dostępne z dowolnego miejsca

3. Oprogramowanie jest regularnie aktualizowane

Zwykle można by powiedzieć, że oprogramowanie jest niedrogie, ale wraz z wiekiem subskrypcji licencje na oprogramowanie nie są już wieczyste, więc miesięczna płatność jest wszystkim, co jest konieczne, aby zapewnić czytelnikowi dostęp do oprogramowania, o ile subskrypcja jest aktualna. Zbadajmy każdy punkt i omówmy go.

Oprogramowanie jest łatwe w użyciu

Jeśli analityk może wybrać kilka przycisków i - voila - pojawia się wynik, jest to znacznie łatwiejsze niż słowo „p". Co to jest słowo „p"? Programowanie! Jeśli analityk musi programować, trudno jest uzyskać wynik. Oczywiście analitycy nie zdają sobie sprawy, że gdy coś jest zaprogramowane, łatwiej jest zastosować to programowanie. Głównym celem jest to, że oprogramowanie z graficznym interfejsem użytkownika (GUI) wydaje się być preferowane względem oprogramowania do programowania. Oprogramowanie COS jest dobrze znane i łatwe w użyciu. Niektóre oprogramowanie FOSS będzie wymagało więcej przygotowań.

Oprogramowanie jest dostępne z dowolnego miejsca

W dobie przetwarzania w chmurze dostęp do oprogramowania wydaje się banalny. Po rozmowie ze współpracownikami podoba im się fakt, że mogą wykonywać i zapisywać swoją pracę online, aby jej nie stracić. Podoba im się również fakt, że aktualizacje są przejrzyste i wykonywane podczas korzystania z narzędzia. Wreszcie podoba im się fakt, że nie muszą martwić się instalacją oprogramowania i wykorzystaniem pamięci lub miejsca na dysku.

Oprogramowanie jest regularnie aktualizowane

W poprzedniej sekcji omówiono to, więc nie będziemy się tym rozwodzić. Należy jednak pamiętać, że narzędzia, które zostaną omówione, są regularnie aktualizowane. Niestety analityk będzie musiał zgodzić się na aktualizacje.

Podsumowanie

Teraz, gdy omówiliśmy, dlaczego analitycy preferują określone narzędzia, opis omawianych tutaj narzędzi zostanie podany w formie tabeli, aby uprościć prezentację i (jak stwierdzono wcześniej) zminimalizować słowo pisane.

Oprogramowanie : Łatwość (1 = łatwe, 5 = trudne) : Dostępność : Aktualizacja

Excel : 1 : 24/7 : Firma

R (RStudio / Rattle) : 3 : 24/7 : Analityk

KNIME : 4 : 24/7 : Analityk

OpenOffice : 2 : 24/7 : Analityk

DLACZEGO ANALIZA DANYCH (DATA SCIENCE) W OGÓLE?

Dzisiejszy świat jest kompendium danych. Dane istnieją we wszystkim, co robimy, niezależnie od tego, czy kupujemy artykuły spożywcze, czy szukamy informacji o zakupie domu. Jest tak wiele bezpłatnych apletów i aplikacji, które są dla nas dostępne, że trudno nam odmówić żadnej z nich. Jak ujął to jeden autor , jeśli to, co pobierasz, jest bezpłatne, to jesteś produktem. To przejmujące, ponieważ darmowe i otwarte oprogramowanie (FOSS) jest czymś powszechnie dostępnym i dostępnym dla nas wszystkich. Jednak dlaczego potrzebujemy nauki o danych do analizy wszystkich tych informacji? W mojej wiedzy istnieje wiele powodów, dla których istnieje nauka o danych. Po pierwsze, istnieje po to, aby zebrać biliony bajtów informacji, które są gromadzone przez firmy i agencje rządowe, aby określić wszystko, od kosztu mleka po ilość emisji dwutlenku węgla do powietrza. Czterdzieści lat temu większość danych była gromadzona, odzyskiwana i przechowywana na papierze. Komputery osobiste były snem, a naukę o danych nazywano archiwizacją lub czymś podobnym. Przechodząc w kierunku mediów elektronicznych, bazy danych zmieniły stosy papieru w kilo-, mega-, gigabajty, a nawet petabajty. Ale przy takiej ilości danych analiza zmieniła się z ołówka i papieru w komputery osobiste lub dowolny komputer. Analitycy zaczęli zdawać sobie sprawę, że dynamiczne oprogramowanie jest sposobem na nadanie analizie danych bardziej użytecznej formy. Nauka o danych wyrosła z tego wysiłku analitycznego i wykorzystuje konwencjonalne metody statystyczne w połączeniu z mocą obliczeniową, aby nauka o danych była łatwo dostępna dla wszystkich podmiotów prywatnych i publicznych. Dzięki możliwości analizy danych marketingowych, technicznych i personalnych firmy mają teraz możliwość obliczania prawdopodobieństwa odniesienia sukcesu przez ich produkt lub wzrostu przychodów w następnym roku. Wraz z rozwojem nauki o danych pojawia się wiele narzędzi, które umożliwiają analizę danych.

GDZIE UZYSKAĆ DANE

Teraz, gdy mamy już wprowadzenie do „dlaczego" nauki o danych, następny temat to „gdzie". Skąd czerpiesz dane do wykorzystania w narzędziach do analizy danych? Odpowiedź na to pytanie, zwłaszcza teraz, brzmi: dane są dostępne do analizy na wielu stronach internetowych. Niektóre z tych witryn internetowych obejmują:

1. www.data.gov, który zawiera strony danych z różnych agencji rządowych. Jeśli chcesz wiedzieć o danych klimatycznych, spisie ludności lub zwalczaniu chorób, to jest miejsce, do którego należy się udać.

2. www.kaggle.com, który nie tylko zawiera dane, ale organizuje konkursy z istniejącymi danymi, do których każdy może dołączyć. Jeden zestaw danych zawiera różne dane zebrane z Titanica, w tym liczbę zgonów lub przeżyć oraz wszystkie dane demograficzne do analizy i korelacji.

3. Prawie każda agencja rządowa. Jeśli nie chcesz wchodzić na ogólną witrynę internetową, przejdź do www.cdc.gov, www.census.gov, ww.noaa.gov lub dowolnej odrębnej rządowej witryny internetowej, aby uzyskać dane dotyczące spraw takich jak ubezpieczenia społeczne (www .ssa.gov) lub nawet informacje wywiadowcze (www.nsa.gov) w przypadku niektórych danych historycznych. Teraz, gdy

wiesz już „dlaczego” i „gdzie” związane z nauką o danych i narzędziami, przechodzisz teraz do następnego kroku, czyli korzystania z narzędzi z prawdziwymi danymi. Poza tym niewątpliwie masz dość tej scenerii. Dane zostały pobrane z witryny :

https://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/,

która zawiera dane śledzenia tornada w Stanach Zjednoczonych od 1951 do 2018 roku. agencja rządowa NOAA oznacza National Oceanic and Atmospheric Agency. Zalecamy pobranie tych plików (tyle, ile chcesz) i używanie ich oddzielnie w przykładach w książce. Ta książka skupi się na śledzeniu tornada z 1951 roku, aby uczynić to stosunkowo prostym. Po pobraniu danych następnym krokiem jest zaimportowanie danych do Twojego ulubionego narzędzia statystycznego.