**PogRomcy danych** to zbiór dwóch kursów o przetwarzaniu, wizualizacji i modelowaniu danych.

Kursy zostały przygotowane dla osób, które pracują, chciałyby pracować, lub są ciekawe jak wygląda, praca z danymi. Oba kursy są prowadzone na poziomie podstawowym. Zakładamy jedynie ogólną orientację jak wygląda tabela z danymi w arkuszach kalkulacyjnych, np. takich jak Excel, oraz ogólną umiejętność pracy z komputerem. Wcześniejsza znajomość programowania będzie bardzo przydatna, ale nie jest konieczna. Wszystkie nowe pojęcia pokazujemy na przykładach związanych z analizą ciekawych danych.

W pierwszym kursie pokażemy jak pracować z programem R, bezpłatnym i bardzo potężnym narzędziem do analizy danych. Pokażemy jak pisać programy, tworzyć funkcje, pętle, wczytywać, przetwarzać i zapisywać dane.

W drugim kursie pokażemy co wyróżnia zły, a co dobry wykres. Następnie pokażemy jak tworzyć dobre wykresy korzystając z programu R. Druga część dotyczy modeli i metod statystycznych. Pokażemy jak badać zależności pomiędzy dwoma lub większą liczbą zmiennych.

Oba kursy zwieńczone są zbiorem 20 zadań do samodzielnego wykonania. Osoby, które wykonają 13 lub więcej zadań otrzymają tytuł ,,Pogromców danych’’ oraz certyfikat ukończenia kursu sygnowany przez Uniwersytet Warszawski.

Aby uzyskać dostęp do materiałów potrzebna jest rejestracja. Jest ona bezpłatna i otwarta dla każdego zainteresowanego. Zrejestrowani uzyskują dostęp do materiałów dydaktycznych, forum oraz zadań.

Materiały dydaktyczne zostały podzielone na krótkie bloki, omawiające kolejne zagadnienia. Każdy blok kończy się krótkimi sprawdzającymi zadaniami. W materiałach znajdują się również przykładowe odpowiedzi do tych zadań.

Do obu kursów dostępne jest forum, ułatwiające komunikacje pomiędzy uczestnikami. Jeżeli jakaś porcja materiału jest trudna, sprawia kłopoty, lub jeżeli w trakcie pracy pojawią się ciekawe rozwiązania, warto podzielić się nimi na forum. Obsługa kursu śledzi forum i odpowiada na wszelkie, nawet najbardziej podstawowe pytania. Również inni uczestnicy kursu mogą coś zasugerować lub pomóc w rozwiązaniu problemu.

Pierwszy kurs, jest poświęcony programowi R. Rozpoczyna się od czterech odcinków pokazujących dlaczego warto nauczyć się tego programu, jak go zainstalować, jak doinstalować wygodny edytor oraz dodatkowe pakiety do pracy z R. Następnie omawiane są podstawy pracy z programem R. Pokażemy jak i gdzie wpisywać komendy oraz gdzie spodziewać się wyników.

Kolejne trzy odcinki poświęcone są wczytywaniu i zapisywaniu tabel z danymi, oraz podstawowym operacjom na tych tabelach z danymi. Pokażemy jak wczytać plik tekstowy lub w formacie Excela do programu R. Jak zapisać wyniki oraz jak z wyników wybrać tylko określone kolumny lub wiersze.

Odcinki 8, 9, 10 pokazują jak automatyzować obliczenia. Wyjaśniamy w nich na przykładach jak działają pętle, instrukcje warunkowe oraz jak tworzyć własne funkcje.

Odcinki od 11 do 15 wprowadzają funkcje i operacje typowe dla najpopularniejszych rodzajów. Pokażemy co można zrobić z wektorem liczb, a co z wektorem napisów lub dat. Wyjaśnimy czym są zmienne logiczne i jakościowe oraz jakie operacje na nich można wykonać.

Odcinki od 16 do 24 przedstawiają szeroki wachlarz funkcji pozwalających na swobodne przekształcanie danych. Opanowawszy te funkcje będziemy potrafili wykonać 90% typowych prac związanych z przetwarzaniem danych. Będziemy wiedzieć jak filtrować, przekształcać, agregować, grupować dane i zmieniać ich strukturę.

Ten kurs jest kursem podstawowym, ale kończy go odcinek pokazujący gdzie można szukać dalszych informacji o programie R. Jeżeli program R przypadnie nam do gustu, w ostatnim odcinku pokażemy jak go dalej zgłębiać.

Pierwszy kurs pokazuje jak przetwarzać dane, drugi kurs pokazuje po co to robić. W tej części przedstawimy zagadnienia związane z eksploracją danych, wizualizacją oraz modelowaniem statystycznym.

Rozpoczynamy od trzech odcinków pokazujących kamienie milowe w historii wizualizacji danych, trudności związane z prezentowaniem danych, problemy wynikające ze sposobu w jaki nasz mózg interpretuje obraz a następnie pokażemy przykłady rzetelnych oraz nierzetelnych wykresów.

Odcinki 4, 5, 6 i 7 wprowadzą nas w świat pakietu ggplot2. Poznamy podstawowe i bardziej zaawansowane instrukcje, pozwalające na wykonanie praktycznie dowolnego wykresu w programie R.

Jeżeli praca z wykresami w programie R nas zaciekawi, to w odcinku 8 przedstawimy sugestie dalszych materiałów rozwijających nasze umiejętności wizualizacji danych.

Kolejne odcinki przedstawią podstawowe narzędzia statystyczne służące do eksploracji danych do analizy statystycznej danych. Pokażemy jak statystycznie porównać dwie średnie. Jak sprawdzić czy dwa zjawiska występują niezależnie czy też czy jest pomiędzy nimi jakaś zależność. Oraz pokażemy jak tworzyć modele regresyjne np. aby prognozować pewną cechę w przyszłości.

Również ten blok zakończy się listą propozycji literaturowych w których można znaleźć więcej informacji na temat modelowania statystycznego.

Do obu kursów dodaliśmy przykłady bardzo ciekawych ale i nietypowych zastosowań analizy danych. Pokażemy jak analiza danych jest wykorzystywana w analizie danych sondażowych, robotyce, inżynierii lingwistycznej i meteorologii.

Analiza danych, a w szczególności analiza z użyciem programu R, charakteryzuje się stromą krzywą uczenia. Na początku wiele rzeczy będzie nowych i trudnych. Przygotowaliśmy jednak wiele materiałów i na bieżąco monitorujemy forum, aby odpowiadać na wszelkie pytania i pomagać w stawianiu pierwszych kroków. Gwarantujemy jednak, że wysiłek włożony w poznawanie programu R opłaci się. Lepiej będziemy mogli zrozumieć co właściwie robimy w analizach. Z czasem będziemy coraz sprawniej przetwarzać dane, a elastyczność i ekspresja programu R powodują, że praktycznie nie będzie przed nami barier związanych z analizą najróżniejszych danych.

Pracując ze slajdami, możemy w każdej chwili nacisnąć przycisk `T`, rozwinie on spis treści lub przycisk `A`, aby zamienić slajdy na ciągły dokument.

Zacznijmy naszą przygodę od krótkiego omówienia programu RStudio. Po pierwszym uruchomieniu, po lewej stronie wyświetli się interaktywna konsola, po prawej stronie są panele pomocnicze. Konsolę otwiera informacja o wersji zainstalowanego programu R, oraz znak zachęty ‘dziubek’. Ten znak, oznacza, że program R jest gotowy do dalszej pracy i możemy wpisywać polecenia.

Zazwyczaj, praca z programem R przebiega w sposób interaktywny.

Jeżeli wpiszemy komendę do konsoli i zatwierdzimy ją klawiszem Enter, to R wykona polecenie i jego wynik wyświetli w konsoli.

Wyniki operacji możemy przypisać do zmiennych. Dzięki temu, zamiast wyświetlać się na ekranie, zostaną one zapamiętane w pamięci programu R, oraz będziemy mogli je wykorzystywać w przyszłości.

Zauważmy, że w oknie po prawej górnej stronie wyświetliły się nowo zdefiniowane zmienne oraz ich wartości. W każdej chwili możemy sprawdzić, jakie zmienne program R aktualnie pamięta, oraz jakie są wartości tych zmiennych.

Prawdziwa siła programu R tkwi w dużym zbiorze funkcji, którymi możemy wykonywać najróżniejsze operacje.

Przypuśćmy, że chcemy zbudować sekwencje liczb od 0 do 10 z krokiem co 0,1.

Operatorem `dwóch znaków zapytania` możemy wyszukać funkcje, która w opisie ma określone słowo lub zwrot. Aby wyświetlić opis dla zadanej funkcji wystarczy wykorzystać operator `znaku zapytania`.

Opis każdej funkcji składa się z opisu jej argumentów, szczegółowego opisu działania funkcji oraz przykładów jej użycia.

Dla funkcji `seq`, aby podać początek, koniec i krok dla sekwencji, należy wskazać argumenty `from`, `to` i `by`.

Argumenty do funkcji możemy podać w dowolnej kolejności.

Mamy już sekwencję liczb. Zróbmy z nią coś ciekawego.

Funkcją `exp` policzmy eksponentę z sekwencji, a następnie - wykorzystując funkcję `plot` - narysujmy i sekwencję, i jej eksponentę.

Pod koniec odcinka przygotowaliśmy zadania do wykonania. Warto zmierzyć się z nimi, gdyż najlepszym sposobem by nauczyć się języka programowania jest ćwiczenie, ćwiczenie i jeszcze raz ćwiczenie. Odpowiedzi do zadań umieszczone są na stronie kursu, można je w każdej chwili sprawdzić.

Zadanie do tego odcinka polega na uruchomieniu wszystkich przedstawionych instrukcji w programie RStudio i sprawdzeniu czy otrzymuje się te same wyniki.