1. **Menjelajahi dan menganalisis data dengan Python**

Penjelajahan dan analisis data adalah inti dari ilmu data. Ilmuwan data memerlukan keterampilan dalam bahasa seperti Python untuk menjelajahi, memvisualisasikan, dan memanipulasi data.

**Tujuan pembelajaran :**

Dalam modul ini, Anda akan mempelajari:

* Tugas penjelajahan dan analisis data umum.
* Cara menggunakan paket Python seperti NumPy, Pandas, dan Matplotlib untuk menganalisis data.

**Prasyarat**

* Pengetahuan tentang matematika dasar
* Beberapa pengalaman pemrograman di Python

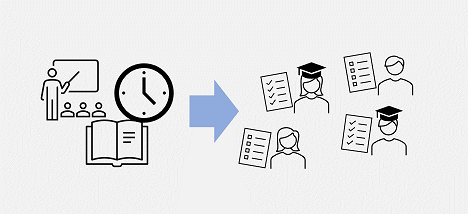
1. **Pengantar**

Tidak mengherankan, peran seorang ilmuwan data sangat terlibat dalam menjelajahi dan menganalisis data. Hasil analisis mungkin menjadi dasar laporan atau model pembelajaran mesin, tetapi semuanya dimulai dengan data, dengan Python menjadi bahasa pemrograman paling populer untuk para ilmuwan data.

Setelah beberapa dekade pengembangan sumber terbuka, Python menyediakan fungsionalitas yang luas dengan pustaka statistik dan numerik yang kuat:

* 1. NumPy dan Pandas menyederhanakan analisis dan manipulasi data
  2. Matplotlib menyediakan visualisasi data yang menarik
  3. Scikit-learn menawarkan analisis data prediktif yang sederhana dan efektif
  4. TensorFlow dan PyTorch menyediakan pembelajaran mesin dan kemampuan pembelajaran mendalam
  5. Biasanya, proyek analisis data dirancang untuk menetapkan wawasan seputar skenario tertentu atau untuk menguji hipotesis.

Misalnya, seorang profesor universitas mengumpulkan data dari mahasiswanya, termasuk jumlah kuliah yang dihadiri, jam yang dihabiskan untuk belajar, dan nilai akhir yang dicapai pada ujian akhir semester. Profesor dapat menganalisis data untuk menentukan apakah ada hubungan antara jumlah belajar yang dilakukan siswa dan nilai akhir yang mereka capai. Profesor mungkin menggunakan data untuk menguji hipotesis bahwa hanya mahasiswa yang belajar selama beberapa jam yang dapat berharap untuk mencapai nilai kelulusan.



1. **Menjelajahi data dengan NumPy dan Pandas**

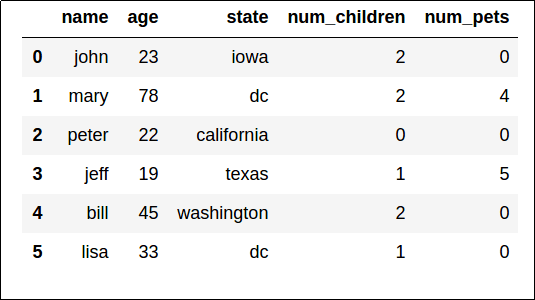
lmuwan data dapat menggunakan berbagai alat dan teknik untuk menjelajahi, memvisualisasikan, dan memanipulasi data. Salah satu cara paling umum ketika ilmuwan data bekerja dengan data adalah dengan menggunakan bahasa Python dan beberapa paket tertentu untuk pemrosesan data.

* 1. Apa itu NumPy?

NumPy adalah pustaka Python yang memberikan fungsionalitas yang sebanding dengan alat matematika seperti MATLAB dan R. Meskipun NumPy menyederhanakan pengalaman pengguna secara signifikan, NumPy juga menawarkan fungsi matematika yang komprehensif.

* 1. Apa itu Pandas

Pandas adalah pustaka Python yang sangat populer untuk analisis dan manipulasi data. Pandas seperti excel untuk Python - menyediakan fungsionalitas yang mudah digunakan untuk tabel data.



**Menjelajahi data di buku catatan Jupyter**

Buku catatan Jupyter adalah cara populer untuk menjalankan skrip dasar menggunakan browser web Anda. Biasanya, buku catatan ini adalah halaman web tunggal, dipecah menjadi bagian teks dan bagian kode yang dijalankan di server, bukan di mesin lokal Anda. Ini berarti Anda dapat memulai dengan cepat tanpa perlu menginstal Python atau alat lain.

**Menguji hipotesis**

Eksplorasi dan analisis data biasanya merupakan proses berulang, di mana ilmuwan data mengambil sampel data dan melakukan jenis tugas berikut untuk menganalisis sampel data dan menguji hipotesis:

* **Bersihkan data** untuk menangani kesalahan, nilai yang hilang, dan masalah lainnya.
* **Terapkan teknik statistik untuk lebih memahami data**, dan bagaimana sampel diharapkan dapat mewakili populasi data dunia nyata, memungkinkan variasi acak.
* **Visualisasikan data** untuk menentukan hubungan antara variabel, dan dalam kasus proyek pembelajaran mesin, identifikasi fitur yang berpotensi memprediksi label.
* Merevisi hipotesis dan mengulangi prosesnya.

1. **Latihan - Menjelajahi data dengan NumPy dan Pandas**

**Exploring Data with Python :** D:\Data Scientist Indonesia - Cloud Skills Challenge from Microsoft\1 - Menjelajahi data dengan NumPy dan Pandas\01-numpy-and-pandas.ipynb

1. **Memvisualisasikan data**

Ilmuwan data memvisualisasikan data untuk memahaminya dengan lebih baik. Tindakan ini bisa berarti melihat data mentah, ukuran ringkasan seperti rata-rata, atau membuat grafik data. Grafik adalah cara yang ampuh untuk melihat data, karena kita dapat membedakan pola yang cukup kompleks dengan cepat tanpa perlu menentukan ukuran ringkasan matematis.

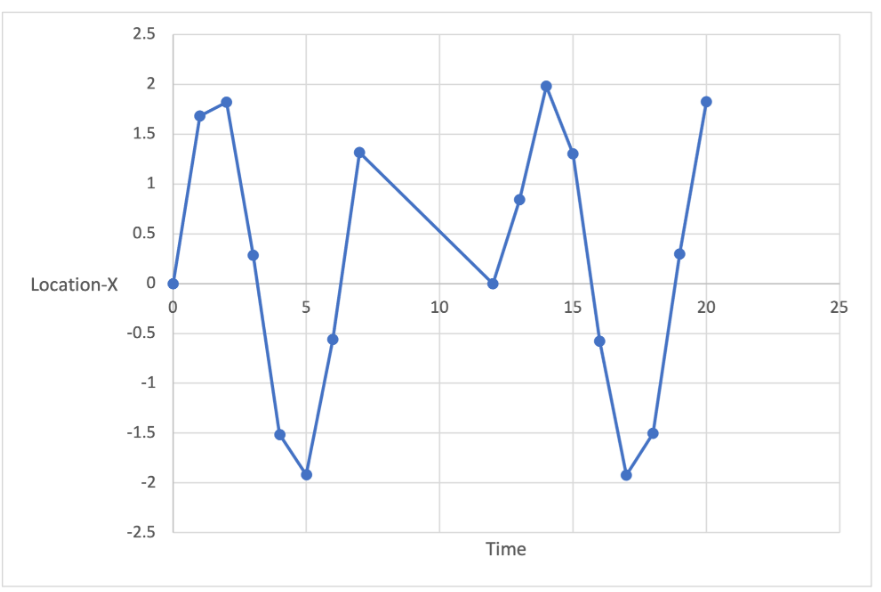
**Mewakili data secara visual**

Mewakili data secara visual biasanya berarti membuat grafiknya. Hal ini dilakukan untuk memberikan penilaian kualitatif cepat dari data kami, yang dapat berguna untuk memahami hasil, menemukan nilai outlier, memahami bagaimana angka didistribusikan, dan sebagainya.

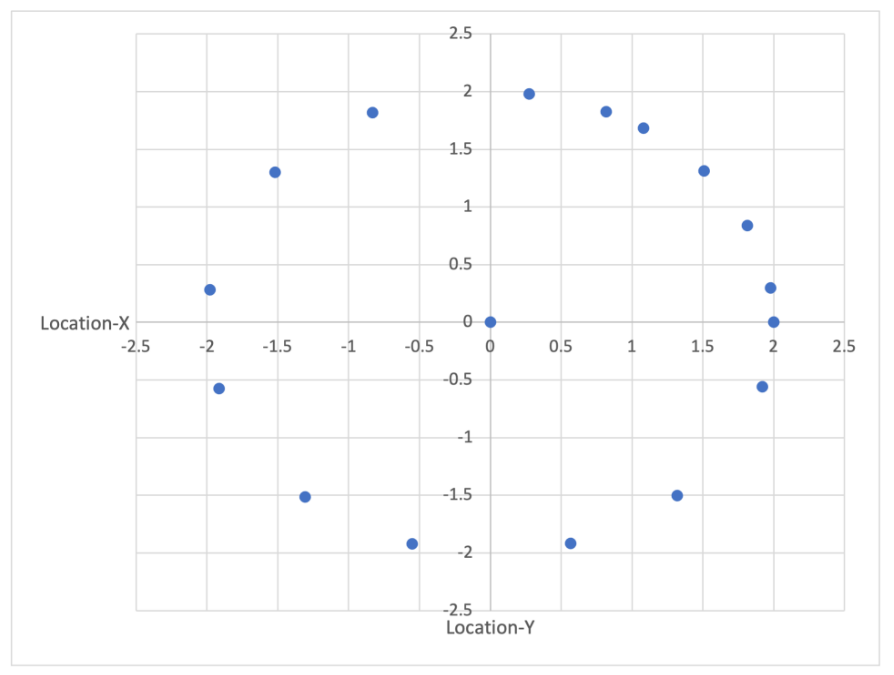
Meskipun terkadang kita tahu sebelumnya jenis grafik apa yang paling berguna, di lain waktu kita akan menggunakan grafik dalam cara yang eksploratif. Untuk memahami kekuatan visualisasi data, perhatikan data di bawah ini: lokasi (x,y) mobil yang mengemudi sendiri. Dalam bentuk mentahnya, sulit untuk melihat pola yang nyata. Mean atau rata-rata, memberi tahu kita bahwa jalurnya berpusat di sekitar x=0.2 dan y=0.3, dan rentang angka tampaknya antara sekitar -2 dan 2.

| **Waktu** | **Lokasi-X** | **Lokasi-Y** |
| --- | --- | --- |
| 0 | 0 | 2 |
| 1 | 1,682942 | 1,080605 |
| 2 | 1,818595 | -0,83229 |
| 3 | 0,28224 | -1,97998 |
| 4 | -1,5136 | -1,30729 |
| 5 | -1,91785 | 0,567324 |
| 6 | -0,55883 | 1,920341 |
| 7 | 1,313973 | 1,507805 |
| 12 | 0,00001 | 0,00001 |
| 13 | 0,840334 | 1,814894 |
| 14 | 1,981215 | 0,273474 |
| 15 | 1,300576 | -1,51938 |
| 16 | -0,57581 | -1,91532 |
| 17 | -1,92279 | -0,55033 |
| 18 | -1,50197 | 1,320633 |
| 19 | 0,299754 | 1,977409 |
| 20 | 1,825891 | 0,816164 |

Jika sekarang kita memplot Lokasi-X dari waktu ke waktu, kita dapat melihat bahwa kita tampaknya memiliki beberapa nilai yang hilang antara waktu 7 dan 12.



Jika kita membuat grafik X vs Y, kita akan mendapatkan peta jalur mobil dikemudikan. Terlihat jelas bahwa mobil telah melaju mengelilingi lingkaran, tetapi pada titik tertentu melaju ke pusat lingkaran tersebut.



Grafik tidak terbatas pada plot sebar 2D seperti di atas, tetapi dapat digunakan untuk menjelajahi tipe data lain, seperti proporsi - ditunjukkan melalui diagram lingkaran, grafik batang bertumpuk - bagaimana data disebarkan - dengan histogram, plot kotak dan garis - dan perbedaan antara dua himpunan data. Seringkali, saat mencoba memahami data atau hasil mentah, kita mungkin bereksperimen dengan berbagai jenis grafik hingga menemukan grafik yang menjelaskan data dengan cara yang intuitif secara visual.

1. **Latihan - Memvisualisasikan data dengan Matplotlib**

**Visualize data :** D:\Data Scientist Indonesia - Cloud Skills Challenge from Microsoft\1 - Menjelajahi data dengan NumPy dan Pandas\02-visualize-data.ipynb

1. **Memeriksa data dunia nyata**

Data yang disajikan dalam materi pendidikan seringkali sangat sempurna, dirancang untuk menunjukkan kepada siswa cara menemukan hubungan yang jelas antara variabel. Data 'dunia nyata' sedikit kurang sederhana.

Karena kerumitan data 'dunia nyata', masalah data mentah harus diperiksa sebelum digunakan.

Dengan demikian, praktik terbaik adalah memeriksa data mentah dan memprosesnya sebelum digunakan, yang mengurangi kesalahan atau masalah, biasanya dengan menghapus titik data yang salah atau memodifikasi data menjadi bentuk yang lebih berguna.

**Masalah data dunia nyata**

Data dunia nyata dapat berisi banyak masalah berbeda yang dapat memengaruhi kegunaan data, dan interpretasi kami terhadap hasilnya.

Penting untuk disadari bahwa sebagian besar data dunia nyata dipengaruhi oleh faktor-faktor yang tidak dicatat pada saat itu. Misalnya, kita mungkin memiliki tabel waktu lintasan mobil balap di samping ukuran mesin, tetapi ada banyak faktor lain yang tidak ditulis—seperti cuaca—mungkin juga berperan. Jika bermasalah, pengaruh faktor-faktor ini sering kali dapat dikurangi dengan meningkatkan ukuran himpunan data.

Dalam situasi lain, titik data yang jelas-jelas di luar perkiraan—juga dikenal sebagai 'outlier'—terkadang dapat dihapus dengan aman dari analisis, meskipun harus berhati-hati untuk tidak menghapus titik data yang memberikan wawasan nyata.

Masalah umum lainnya dalam data dunia nyata adalah bias. Bias mengacu pada kecenderungan untuk memilih jenis nilai tertentu lebih sering daripada yang lain, dengan cara yang salah menggambarkan populasi yang mendasarinya, atau 'dunia nyata'. Bias terkadang dapat diidentifikasi dengan menjelajahi data sambil mengingat pengetahuan dasar tentang dari mana data tersebut berasal.

Ingat, data dunia nyata akan selalu memiliki masalah, tetapi hal ini sering kali merupakan masalah yang dapat diatasi. Ingatlah untuk:

1. Memeriksa nilai yang hilang dan data yang direkam dengan buruk
2. Mempertimbangkan penghapusan outlier yang jelas
3. Mempertimbangkan faktor dunia nyata apa yang mungkin memengaruhi analisis Anda dan mempertimbangkan apakah ukuran himpunan data Anda cukup besar untuk menangani ini
4. Memeriksa data mentah yang bias dan pertimbangkan opsi Anda untuk memperbaikinya, jika ditemukan
5. **Exploring data with Python - real world data**

**Explore real world data:** D:\Data Scientist Indonesia - Cloud Skills Challenge from Microsoft\1 - Menjelajahi data dengan NumPy dan Pandas\03-real-world-data.ipynb

1. **Ringkasan**

Dalam modul ini, Anda mempelajari cara menggunakan Python untuk menjelajahi, memvisualisasikan, dan memanipulasi data. Penjelajahan data adalah inti dari ilmu data, dan merupakan elemen penting dalam analisis data dan pembelajaran mesin.

Pembelajaran mesin adalah subset ilmu data yang berkaitan dengan pemodelan prediktif. Dengan kata lain, pembelajaran mesin menggunakan data untuk membuat model prediktif, untuk memprediksi nilai yang tidak diketahui. Anda dapat menggunakan pembelajaran mesin untuk memprediksi berapa banyak makanan yang perlu dipesan supermarket, atau untuk mengidentifikasi tanaman dalam foto.

Pembelajaran mesin bekerja dengan mengidentifikasi hubungan antara nilai data yang menggambarkan karakteristik sesuatu—fitur-fiturnya, seperti tinggi dan warna tanaman—dan nilai yang ingin kita prediksi—label, seperti jenis tumbuhan. Hubungan ini dibangun menjadi model melalui proses pelatihan.

**Tantangan: Menganalisis Data Penerbangan**

Jika latihan dalam modul ini telah menginspirasi Anda untuk mencoba menjelajahi data sendiri, mengapa tidak mengambil tantangan himpunan data dunia nyata yang berisi catatan penerbangan dari Departemen Transportasi Amerika Serikat? Anda akan menemukan tantangannya di buku catatan 01 - Flights Challenge.ipynb!

<https://github.com/MicrosoftDocs/ml-basics/blob/master/challenges/01%20-%20Flights%20Challenge.ipynb>

