Pendahuluan

Pada bagian sebelumnya, telah dirancang strategi implementasi *big data analytics* untuk perusahaan retail TagMedia yang menjual produk berupa *office & school supply*. Produk tersebut terbagi menjadi beberapa kategori seperti kertas, binder, amplop, barang elektronik, dan furnitur perkantoran. Proses bisnis yang dilaksanakan dalam TagMedia dimulai dari pembelian produk dari berbagai *supplier* dengan merk dagang dan jenis produk yang berbeda-beda. Produk tersebut umumnya didapatkan dengan harga yang lebih murah karena dibeli langsung dari tangan pertama. TagMedia kemudian menaikkan harga dari produk dan menjualnya kepada *customer*, selisih kenaikan harga tersebut menjadi keuntungan bagi TagMedia. Pelanggan dari TagMedia tidak hanya *customer* perorangan, tetapi juga perusahaan atau perkantoran lain untuk digunakan sendiri (tidak dijual kembali).

TagMedia memiliki tiga *offline store* yang terletak di California, Michigan, dan Florida. Selain itu, TagMedia juga menyediakan produknya untuk dibeli secara *online* melalui *website* TagMedia.com dan *e-commerce* seperti Amazon dan e-Bay. TagMedia memiliki data dari berbagai sumber, yaitu sistem kasir, gudang, hingga *website* untuk merekam data-data penjualan dan produk. Data tersebut tersimpan dalam *data lake* dan *data warehouse* sesuai dengan rancangan pada proyek semester sebelumnya. Dalam rangka memenangkan persaingan bisnis, TagMedia perlu mengubah data-datanya menjadi *smart data* melalui proses *data science*. Proses tersebut terdiri atas beberapa tahap, yaitu *business data understanding*, *data preparation*, *modelling*, *deployment*, dan menghasilkan *outcome* berupa *knowledge* yang membantu perusahaan untuk menyusun strategi bisnis dengan tepat.

Eksplorasi Data

Proses pertama dalam tahapan *data science* adalah pengenalan terhadap data dan bisnis. Secara singkat, TagMedia mengalami permasalahan yaitu bisnis dijalankan tanpa landasan yang kuat, TagMedia beberapa kali mengalami kerugian akibat membeli barang dengan *demand* yang rendah dan menjual barang di bawah harga yang seharusnya. Namun, TagMedia ingin mengembangkan bisnisnya agar semakin besar, menjangkau pasar yang lebih luas, mendapatkan lebih banyak keuntungan, dan meminimalisir terjadinya kerugian. Data yang dimiliki oleh TagMedia berupa data transaksi, pelanggan, dan produk. Pengenalan mengenai data dan bisnis secara lebih detail telah dijelaskan pada proyek perancangan strategi implementasi *big data analytics* di semester sebelumnya. Pada bagian ini, proses *data science* memasuki tahapan kedua, yaitu *data preparation*.

Data preparation, atau persiapan data, merupakan tahapan yang sangat penting dalam proses *data science*. Dalam *data science*, data digunakan sebagai *input* yang akan diolah dengan algoritma sehingga menghasilkan *output* berupa pengetahuan atau *insight*. Apabila data yang digunakan masih berupa data yang tidak terstruktur, berantakan, dan terdiri atas variabel atau *value* yang tidak berarti, maka *outcome* yang didapat juga tidak akan maksimal (*garbage in, garbage out*). Sebagian besar algoritma *data science* membutuhkan data yang terstruktur untuk mendukung performa, dengan demikian dibutuhkan proses persiapan data yang baik. *Data preparation* terdiri atas beberapa bagian, salah satunya adalah *data exploration*. *Data exploration* merupakan tahapan yang dilakukan untuk memahami dan menjelajahi struktur dari data, mengetahui distribusi, *outlier*, dan relasi dari data. Eksplorasi ini bertujuan untuk

memahami data dengan lebih baik, mempersiapkan data untuk analisis yang lebih canggih, dan mendapatkan pengetahuan lebih cepat. *Data exploration* terbagi menjadi dua tipe, yaitu *descriptive statistic* dan *data visualization*. Masing-masing tipe akan dijabarkan untuk menjawab soal ini.

A. *Data exploration* dengan *descriptive statistic*

Statistika deskriptif merupakan tahapan yang sangat umum dilakukan dalam proses *data science*. Statistika deskriptif dilakukan untuk mengetahui karakteristik dari data dengan mengubahnya menjadi metrik numerik sederhana. Berdasarkan banyaknya atribut yang digunakan, statistika deskriptif terbagi menjadi dua jenis, yaitu *univariate* dan *multivariate*. Statistika deskriptif *univariate* merupakan perhitungan yang dilakukan untuk masing-masing satu atribut, tipe *univariate* dilakukan untuk mengetahui karakteristik data seperti ukuran pemusatan data dan persebaran data. Sedangkan *multivariate* menganalisis lebih dari satu atribut sekaligus untuk mengetahui korelasi antar variabel.

Statistika deskriptif *univariate* dan *multivariate* diterapkan pada data TagMedia yang berjenis numerik. Data tersebut meliputi variabel *age, sale, discount, quantity, profit*, dan *shipment duration*. Ukuran pemusatan data terdiri atas *mean* (nilai rata-rata) dan median (nilai tengah), sedangkan sebaran data terdiri atas jangkauan, standar deviasi, dan varians. Jangkauan merupakan selisih data terbesar dengan data terkecil dalam dataset, sedangkan standar deviasi dan varians digunakan untuk mengetahui sebaran data terhadap nilai rata-rata. Standar deviasi dan varians dapat mengetahui seberapa dekat atau seberapa jauh data dari nilai rata-rata. Semakin besar nilai standar deviasi dan varians, dapat disimpulkan bahwa data semakin tersebar jauh dari *mean.* Persebaran data ini dilihat untuk mengetahui perbedaan setiap nilai terhadap nilai rata-rata sehingga menambah pemahaman atas data yang dimiliki. Di dalam pelaksanaannya, pertama-tama setiap data numerik TagMedia diproses dengan oprerator agregat untuk mencari seluruh nilai *univariate*. Proses pencarian ini dilakukan melalui *software* RapidMiner terhadap variabel-variabel berikut:

- *Age*

  Variabel *age* merepresentasikan usia dari pelanggan TagMedia yang melakukan transaksi. Variabel ini berguna untuk mengetahui pembagian pelanggan berdasarkan usianya. *Age* memiliki tipe data numerik, dengan demikian dapat dicari *mean*, median, modus, dan lain-lain dengan hasil sebagai berikut. *Summary* di bawah ini dihasilkan melalui perangkat lunak RapidMiner:

Row No.	average(Age)	median(Age)	standard_d...	variance(Age)	maximum(A...	minimum(A...	mode(Age)
1	26.856	23	12.493	156.064	60	12	25

  Pada variabel ini, rata-rata usia pembeli adalah 26 tahun dengan nilai tengah yang sama yaitu 23 tahun, usia ini didominasi oleh pekerja yang membutuhkan alat-alat kantor. Nilai yang paling sering muncul dalam data ini ada 25, dengan demikian pembeli terbanyak berada pada usia 25 tahun. Nilai maksimal pada data usia adalah 60 tahun, sedangkan data minimal berada pada 12 tahun. Melalui hasil tersebut, diketahui bahwa jangkauan pelanggan TagMedia berada pada rentang 12-60 tahun. Hasil standar deviasi adalah 12,49 maka rata-rata perbedaan setiap poin data terhadap *mean* adalah 12, sedangkan varians pada variabel usia sebesar 156,064.

- *Sales*

Variabel *sales* merupakan variabel yang menampung nilai jual dari suatu poduk. Berkut ini merupakan *summary* untuk variabel *sale*:

average(Sal...	median(Sale...	standard_d...	variance(Sal...	maximum(S...	minimum(S...	mode(Sales)
231.687	55.424	624.398	389872.794	22638.480	0.444	12.960

Pada variabel ini, rata-rata penjualan sebesar $231,69 dengan nilai tengah yang sama yaitu $55,42. Nilai maksimal penjualan mencapai $22638,48 dalam satu kali transaksi, sedangkan pembelian minimal $0,44. Jangkauan nilai *sales* adalah $22638,036. Jangkauan ini tergolong sangat jauh antara nilai minimal dan maksimal. Nilai modus (*mode*) berada pasa nilai $12,96 yang berarti bahwa mayoritas pembelian berada pada nilai 12 dollar. Hasil standar deviasi adalah $624,40, maka rata-rata perbedaan setiap poin data terhadap *mean* adalah $624,40, sedangkan varians pada variabel usia sebesar 389872,794. Standar deviasi dan varians ini tergolong tinggi yang berarti data memiliki sebaran yang jauh.

- *Discount*

  Variabel *discount* merupakan variabel yang menampung nilai diskon atau potongan harga terhadap produk yang dibeli oleh pelanggan. Berikut ini merupakan *summary* untuk variabel diskon:

average(Dis...	median(Disc...	standard_d...	variance(Dis...	maximum(D...	minimum(Di...	mode(Disco...
0.157	0.200	0.207	0.043	0.800	0	0

  Pada variabel ini, nilai rata-rata diskon yang diberikan untuk pelanggan pada setiap produk adalah $0,157 dengan nilai median $0,2. Standar deviasi sebesar $0,207 dan varians sebesar $0,043. Nilai potongan harga maksimal yang diberikan adalah $0,8 dan nilai minimalnya adalah 0 (tidak memberikan diskon). Modus pada variabel ini adalah 0, artinya TagMe lebih banyak tidak memberikan potongan harga kepada pelanggannya.

- *Quantity*

  Variabel *quantity* merupakan variabel yang menampung nilai kuantitas pembelian produk pada TagMedia. Berikut ini merupakan *summary* dari variabel kuantitas:

average(Qu...	median(Qua...	standard_d...	variance(Qu...	maximum(Q...	minimum(Q...	mode(Quant...
3.796	3	2.226	4.954	14	1	3

  Pada variabel ini, nilai rata-rata kuantitas pembelian untuk setiap produk adalah 3 dengan median dan modus sejumlah 3. Hal ini menunjukkan bahwa mayoritas pelanggan membeli 3 unit produk. Sebaran kuantitas produk dapat dilihat melalui nilai maksimal pembelian yaitu 14, serta nila minimal yaitu 1. Jangkauan dalam data ini bernilai 13, artinya pelanggan membeli produk yang berkisar antara 1-14 unit. Standar deviasi memiliki angka 2,226 dengan varians 4,954, kedua nilai ini tergolong kecil, sehingga sebaran data umumnya tidak terlalu jauh.

- *Profit*

  Variabel *profit* merupakan variabel yang menampung nilai keuntungan TagMedia dari hasil penjualan produk. Berikut ini merupakan *summary* dari variabel *profit*:

average(Pro...	median(Prof...	standard_d...	variance(Pr...	maximum(P...	minimum(Pr...	mode(Profit)
28.333	8.570	235.006	55227.824	8399.980	-6600.680	6.220

Pada variabel ini, nilai rata-rata keuntungan yang didapatkan oleh TagMedia adalah $28,333 dengan median sebesar $8,570. *Profit* tersebut merupakan keuntungan bagi TagMedia dengan menaikkan harga jual produk dari harga asli. Nilai *profit* maksimal yang didapatkan selama transaksi tersebut mencapai nilai $8399,980 dengan nilai minimal *profit* adalah -$6600,680. Nilai negatif tersebut tidak dapat dikatakan sebagai keuntungan, melainkan kerugian yang didapatkan oleh TagMedia. Dalam kasus tersebut, TagMedia menjual produk jauh di bawah harga modal sehingga terjadi kerugian. Rentang nilai variabel ini berkisar antara -6600,680 hingga 8399,976 dengan nilai modus 6,22. Nilai ini berarti bahwa TagMedia mendapatkan *profit* yang paling sering di angka 6,22 dollar untuk tiap transaksi. Sebaran data berupa standar deviasi menyentuh nilai 235,006 dengan varians 55227,824, artinya distribusi data tersebar jauh dari rata-rata.

- *Ship duration*
  Variabel *ship duration* merupakan variabel yang menampung nilai durasi atau lamanya suatu barang dikirim hingga sampai kepada pelanggan. Berikut merupakan *summary* dari data tersebut:

average(Shi...	median(Ship...	standard_d...	variance(Shi...	maximum(S...	minimum(S...	mode(ShipD...
3.958	4	1.744	3.040	7	0	4

Pada variabel ini, nilai rata-rata durasi pengiriman adalah 3 hingga 4 hari dengan median dan modus bernilai 4. Mayoritas pengiriman produk membutuhkan waktu sekitar 4 hari, namun hal ini bisa beragam tergantung dengan jenis pengiriman yang dipilih oleh pelanggan. Nilai maksimum pada data ini adalah 7 hari dan nilai minimal 0 hari, jangkauan lamanya pengiriman berkisar antara 0-7 hari. Nilai standar deviasi sebesar 1,744 dengan varians sebesar 3,040. Berdasarkan hasil tersebut, sebaran data *ship duration* tergolong tidak terlalu jauh dari nilai rata-rata.

- Data kategorikal lain
  Data kategorikal lain tidak dapat melalui perhitungan seperti di atas, namun dapat dilihat sebaran frekuensi data untuk setiap kategori yang ada dalam variabel seperti di bawah ini. Secara lebih detail, setiap variabel ini akan dideskripsikan melalui metode visualisasi. Data-data kategorikal yang terdapat pada *data set* yaitu *ShipMode* yaitu metode pengiriman barang, *gender* sebagai jenis kelamin pelanggan, *category* dan *sub category* sebagai pembagian kategori produk yang dijual oleh TagMedia, *segment* sebagai tipe pelanggan pada TagMedia, *orderday* sebagai hari dimana penjualan terjadi, is_weekend sebagai penanda apakah transaksi terjadi di *weekend* atau *weekdays*, is_profitable sebagai penanda apakah transaksi tersebut membawa kerugian atau keuntungan, dan variabel geografis seperti kota, negara bagian, dan *region*.

ShipMode	Polynominal	0	Least: Same Day (349)	Most: Second Class (4925)
Gender	Polynominal	0	Least: Female (4010)	Most: Male (5940)
Category	Polynominal	0	Least: Technology (1860)	Most: Office Supplies (5963)
SubCategory	Polynominal	0	Least: Copiers (69)	Most: Binders (1529)
Segment	Polynominal	0	Least: Corporate (1695)	Most: Consumer (6144)
Country	Polynominal	113	Least: United States (9950)	Most: United States (9950)
City	Polynominal	113	Least: Yucaipa (1)	Most: New York City (905)
State	Polynominal	113	Least: Wyoming (1)	Most: California (1984)
Region	Polynominal	113	Least: South (1619)	Most: West (3180)
OrderDay	Polynominal	113	Least: Wednesday (376)	Most: Monday (1846)
Is_weekend	Binominal	113	Negative: 1	Positive: 0
Is_profitable	Binominal	113	Negative: 1	Positive: 0

Pada bagian analisis *multivariate*, dilakukan pengecekan korelasi atau hubungan antara satu variabel dengan variabel lainnya dalam *dataset*. Pengujian korelasi dilakukan dengan metode Pearson menggunakan variabel numerik. Korelasi memiliki rentang nilai -1 hingga 1, apabila korelasi mendekati 0 maka hubungan antar variabel semakin lemah. Apabila nilai korelasi mendekati atau sama dengan 1, maka kedua variabel tersebut saling mempengaruhi secara kuat. Nilai korelasi negatif menandakan hubungan yang berbanding terbaik, sedangkan nilai positif menandakan hubungan yang searah antar variabel terkait.

Berikut ini merupakan korelasi setiap variabel numerik dalam *dataset*. Hubungan yang dimiliki antar variabel di bawah ini berada pada rentang 0 hingga 0,922. Terdapat hubungan yang beragam pada pengecekan ini, beberapa variabel memiliki hubungan yang sangat kuat hingga sedang. Variabel yang memiliki hubungan kuat adalah antara *price each* dengan *sales*, *price each* dengan *origin price*, dan *origin price* dengan *sales*. Hubungan dengan kekuatan sedang dimiliki oleh *profit* dengan *sales*, *profit* dengan *price each*, *age* dengan *price each*, *age* dengan *sales*, dan *age* dengan *origin price*. Secara umum, penjualan dalam TagMedia dipengaruhi oleh usia pembeli, kuantitas barang, harga barang. Semakin tinggi usia, maka lebih tinggi kemungkinan untuk membeli barang dengan harga yang tinggi. Semakin tinggi harga modal, maka akan semakin tinggi harga jual dan *sales* yang didapatkan. Semakin tinggi nilai *sales*, maka *profit* juga akan semakin tinggi (kekuatan dan kemungkinan sedang).

Attribut...	Age	Quantity	Priceea...	Sales	Discount	Originpr...	Profit	ShipDur...
Age	1	0.212	0.463	0.443	-0.059	0.476	0.128	-0.013
Quantity	0.212	1	-0.003	0.202	0.012	0.000	0.065	0.018
Priceeach	0.463	-0.003	1	0.889	-0.034	0.922	0.466	-0.013
Sales	0.443	0.202	0.889	1	-0.028	0.835	0.476	-0.007
Discount	-0.059	0.012	-0.034	-0.028	1	0.068	-0.222	-0.001
Originpri...	0.476	0.000	0.922	0.835	0.068	1	0.126	-0.014
Profit	0.128	0.065	0.466	0.476	-0.222	0.126	1	-0.005
ShipDur...	-0.013	0.018	-0.013	-0.007	-0.001	-0.014	-0.005	1

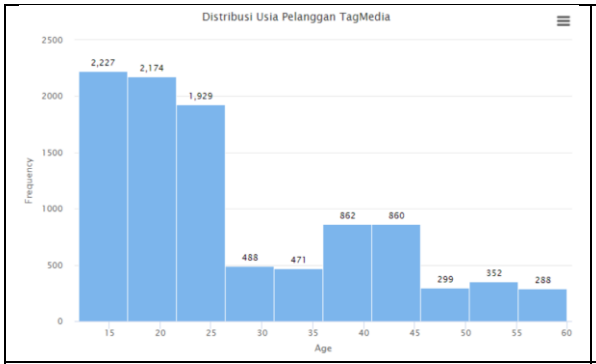
B. *Data exploration* dengan visualisasi

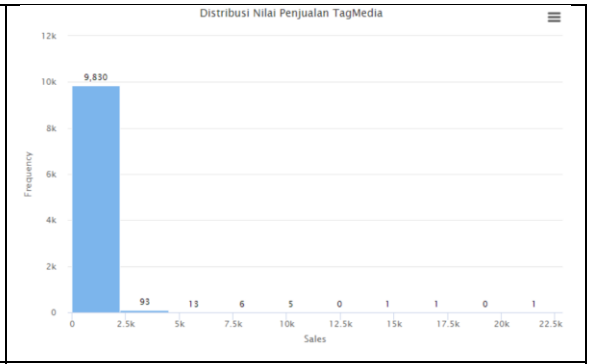
Proses eksplorasi data tidak hanya dilakukan melalui statistika deskriptif, tetapi juga dapat dilakukan melalui visualisasi. Visualisasi merupakan teknik untuk merepresentasikan data dalam bentuk diagram, animasi, gambar, dan bentuk visual lainnya untuk menyediakan informasi sehingga lebih mudah dipahami.

- Distribusi variabel numerik

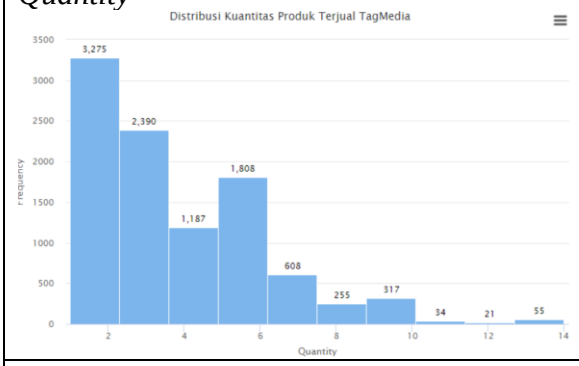
  Apabila dilihat dari distribusinya, seluruh data tidak mengikuti distribusi normal dan tidak simetris. Jika diliat melalui setiap grafik, *skewness* yang dimiliki mayoritas bernilai positif (*right skewed*) atau memiliki *tail* sebelah kanan yang lebih panjang. Distribusi ini dibuat dengan menggunakan histogram:

Age	*Sales*

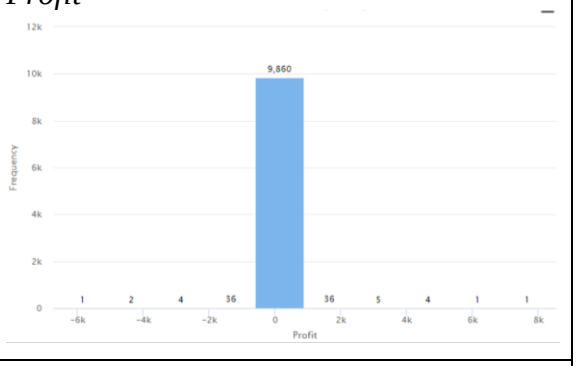




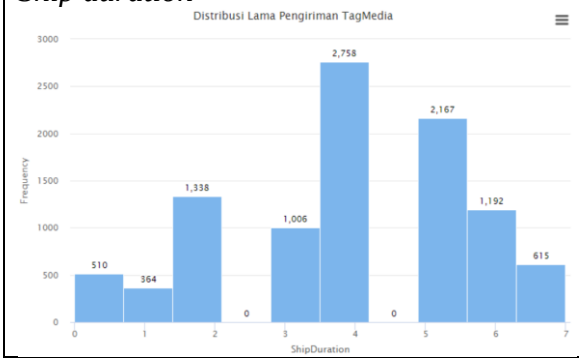
Quantity



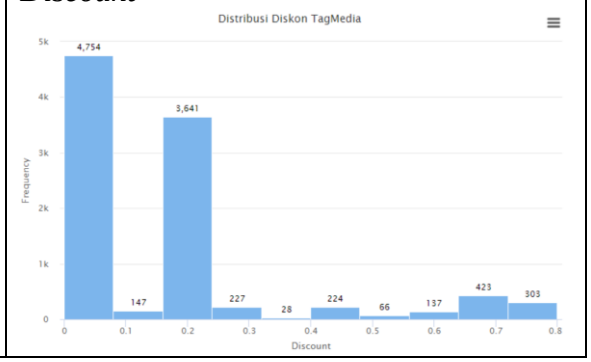
Profit



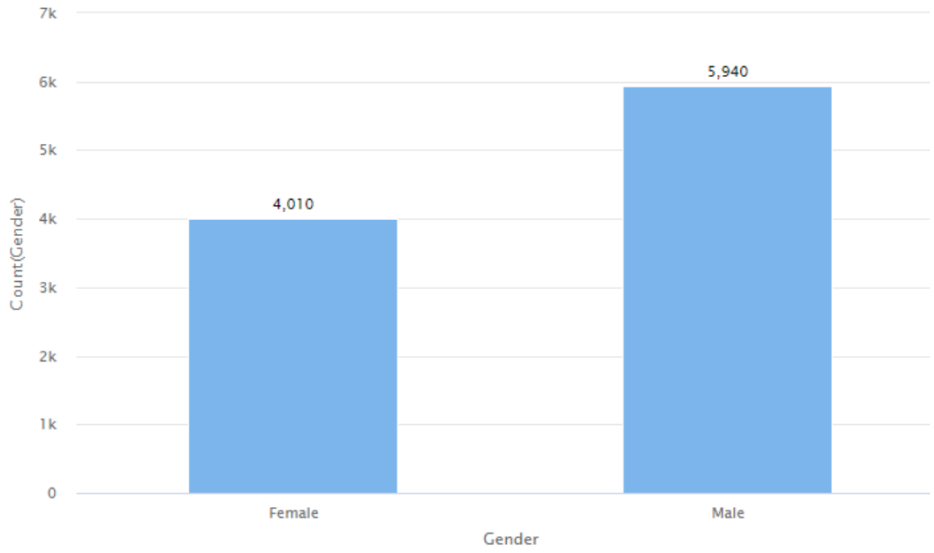
Ship duration



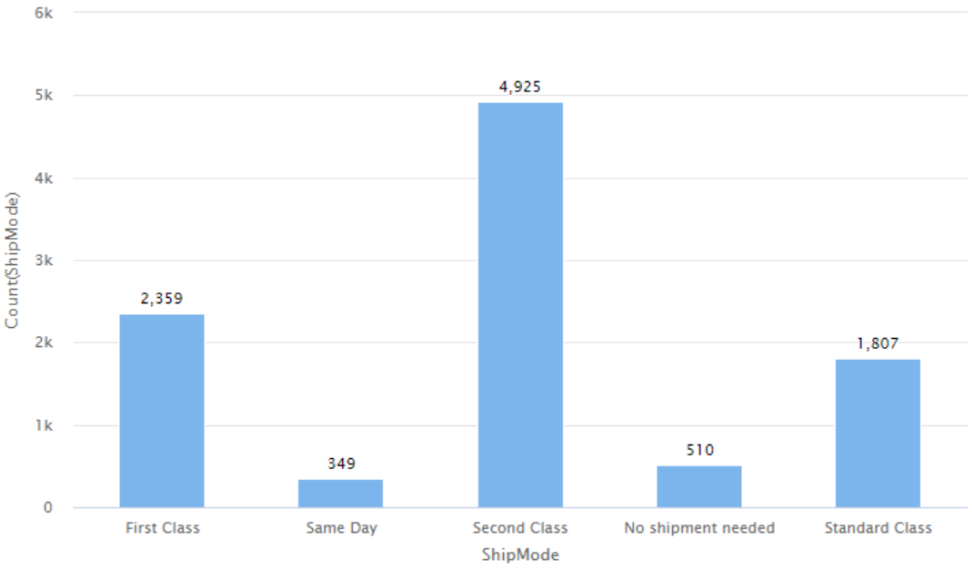
Discount



- Distribusi jenis kelamin pelanggan TagMedia
  Dalam *dataset*, terdapat variabel *gender* yang menampung nilai jenis kelamin pelanggan TagMedia. Berikut merupakan distribusi berdasarkan jenis kelamin. Melalui diagram batang ini, diketahui bahwa jumlah pelanggan laki-laki mencapai 6000 pelanggan, sedangkan wanita berjumlah 4000. Dapat disimpulkan bahwa mayoritas pelanggan TagMedia berjenis kelamin laki-laki.

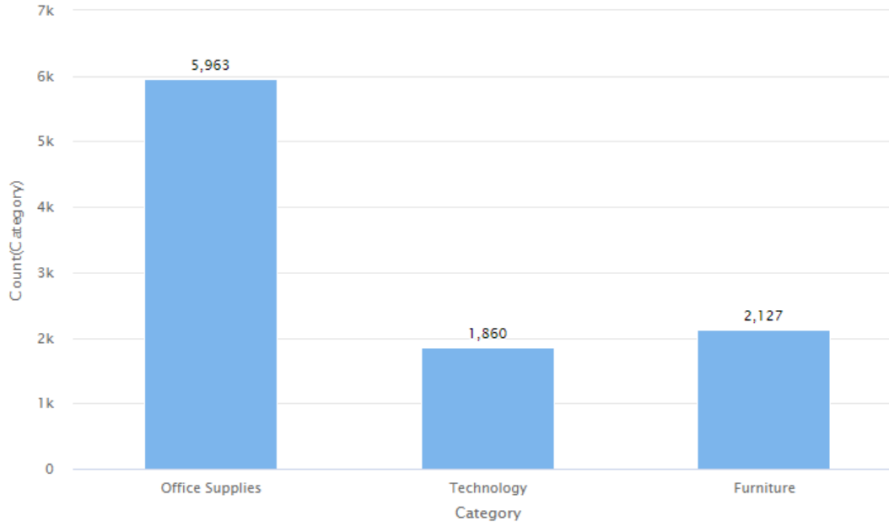


- Distribusi jenis pengiriman TagMedia
  TagMedia memiliki jenis jasa pengiriman yang berbeda-beda, yaitu *same day*, *first class*, *second class*, *standard*, dan *no shipment needed*. *Same day* merupakan layanan yang menjamin bahwa barang akan sampai pada hari yang sama setelah dikirimkan. *First class* merupakan layanan di bawah *same day* yang memiliki waktu pengiriman berkisar antara 2-3 hari. *Second class* memiliki waktu pengiriman 4-5 hari, sedangkan *standard class* umumnya mengirimkan barang dengan durasi 4-7 hari. *No shipment needed* memiliki arti bahwa transaksi pembelian tersebut tidak membutuhkan layanan pengiriman, pelanggan langsung mendatangi *store* TagMedia untuk berbelanja dan membawa produknya sendiri. Melalui diagram di bawah ini, jenis layanan yang paling banyak digunakan adalah *second class*, disusul oleh tipe *first class* dan *standard class*.

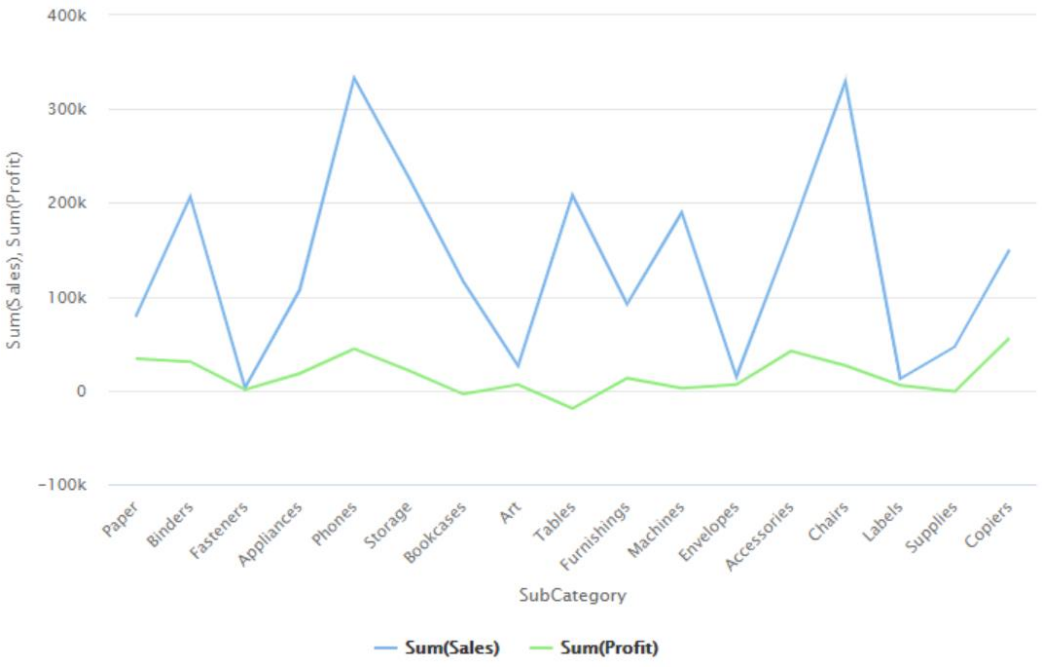


- Distribusi kategori produk berdasarkan penjualan

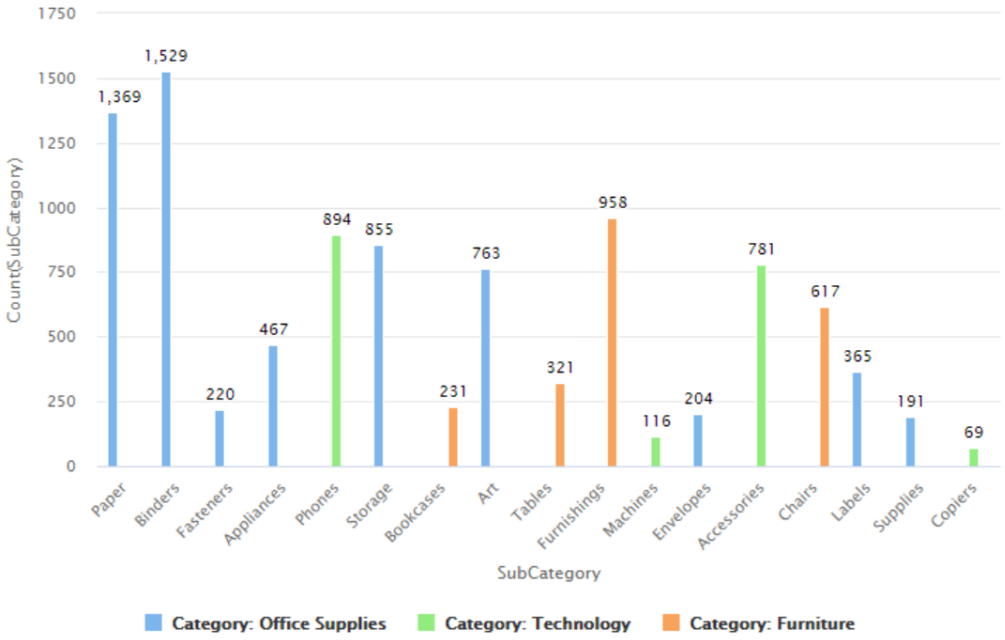
TagMedia menyediakan berbagai jenis produk yang dikelompokkan ke dalam 3 kategori besar, yaitu *office supplies*, *technology*, dan *furniture*. Ketiga produk tersebut masih berhubungan dengan kebutuhan kantor dan sekolah. Melalui diagram di bawah ini, dapat disimpulkan bahwa kategori *office supplies* memiliki jumlah transaksi yang paling banyak dibandingkan kategori lain, yaitu mencapai angka 6000.



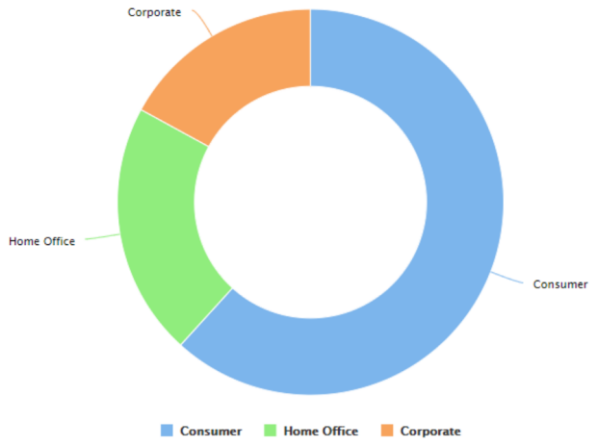
- Penjualan dan keuntungan berdasarkan sub-kategori produk
  Produk dalam TagMedia juga dibagi ke dalam beberapa sub-kategori yang lebih kecil, diagram garis di bawah ini menunjukkan besarnya jumlah *sales* dan *profit* yang didapatkan dari penjualan sub-kategori produk TagMedia. Penjualan tertinggi ada pada sub-kategori *phone* dan *accessories* dengan nilai mencapai $300,000. Sedangkan keuntungan tertinggi ada pada sub-kategori *phone*, *accessories*, dan *copiers* dengan nilai mencapai $55,000.



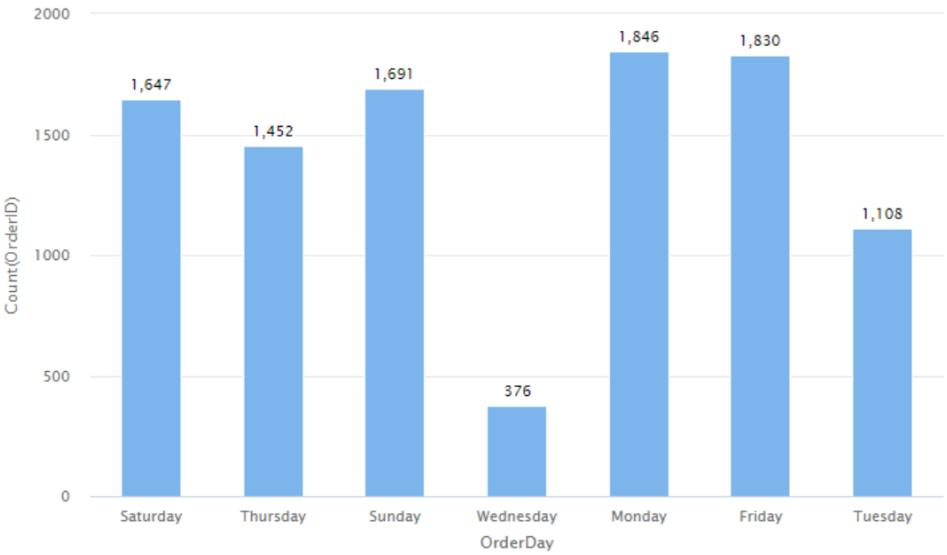
- Jumlah penjualan berdasarkan sub-kategori produk
  Jumlah penjualan sub-kategori produk dapat dilihat berdasarkan diagram di bawah ini. Jumlah penjualan terbanyak berada pada sub-kategori binder dan kertas, sedangkan sub-produk yang paling sedikit peminat adalah *copiers*, *fasteners, machines*, dan *bookcases*. Apabila dibandingkan dengan diagram di atas, meskipun penjualan *copiers* paling sedikit, tetapi dapat menghasilkan *profit* yang besar.



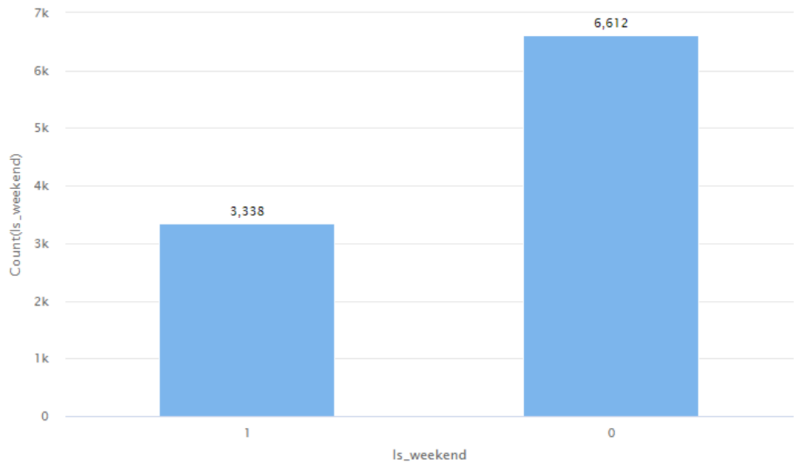
- Distribusi tipe pelanggan TagMedia
  Berdasarkan data yang dimiliki, TagMedia membagi jenis pelanggan ke dalam tiga tipe, yaitu *consumer, corporate,* dan *home office*. *Customer* merupakan pembeli untuk penggunaan perseorangan (*regular customer*), *corporate* merupakan pelanggan yang membeli produk untuk kebutuhan pengunaan pada kantor atau perusahaan, sedangkan *home office* merupakan tipe pelanggan yang menggunakan alat-alat tersebut untuk kebutuhan pekerjaan di rumah. Diagram berikut ini menunjukkan distribusi frekuensi untuk ketiga tipe pelanggan yang ada. Diketahui bahwa mayoritas pelanggan berasal dari tipe *consumer* dengan persentase lebih dari 50% atau 6000 pelanggan, disusul dengan *home office* sebesar 2100 pelanggan, dan *corporate* sebesar 1600 pelanggan.



- Jumlah pembelian berdasarkan hari
  Pada *dataset* ini terdapat nama-nama hari saat transaksi berlangsung, berikut ini merupakan distribusi penjualan berdasarkan hari untuk mengetahui hari apa yang memiliki lebih banyak transaksi. Selain itu juga dilihat distribusi berdasarkan *weekend* dan *weekdays*. Hasilnya berada pada dua diagram di bawah ini, penjualan yang paling banyak terjadi pada hari senin dan jumat, sedangkan hari rabu memiliki penjualan yang sangat rendah dibandingkan hari-hari lainnya. TagMedia dapat mengeksplorasi lebih lanjut mengenai penyebab turunnya penjualan di hari tersebut.

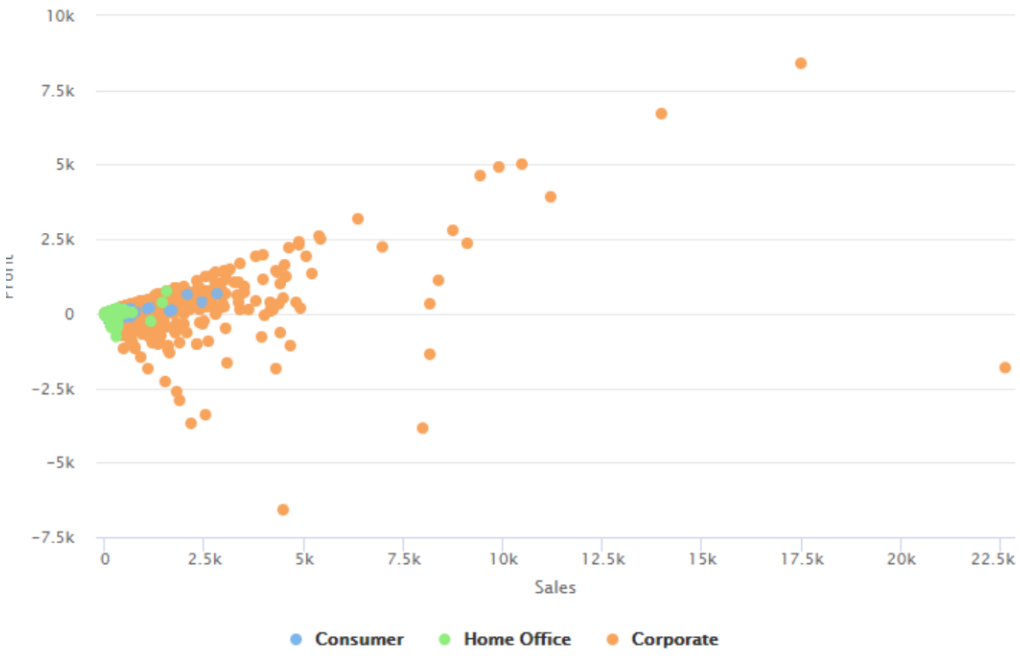


  Selain itu, berdasarkan diagram didapatkan informasi bahwa pembelian alat-alat kantor lebih sering terjadi pada *weekdays* dibandingkan *weekend*. Secara umum, hari libur justru memiliki jumlah penjualan yang lebih sedikit, meskipun pada *weekdays* di hari rabu penjualannya tergolong rendah.



- Hubungan *profit* dengan *sales*

Berikut ini merupakan diagram yang menjelaskan hubungan antara variabel keuntungan dengan penjualan. Secara umum, diagram ini menunjukkan hubungan linear antara kedua variabel, namun tidak dalam bentuk yang jelas dan tegas. Data masih berkumpul dalam satu titik sehingga memperkecil hubungan linear antara kedua variabel. Hal ini menjadi salah satu validasi terhadap masalah yang dimiliki oleh TagMedia. TagMedia melaksanakan bisnisnya tanpa melakukan perhitungan yang matang atas *profit* dan *loss*, sehingga penjualan harga barang dengan *profit* yang akan didapatkan cenderung tidak bersifat linear (hanya mempengaruhi dalam skala kekuatan sedang). Melalui diagram ini, dapat dilihat bahwa nilai *profit mayoritas* berkumpul mendekati titik 0 meskipun *sales* meningkat. Bahkan beberapa titik data berada pada *profit* negatif (*loss*) dalam diagram tersebut. Berdasarkan hal tersebut, TagMedia perlu membenahi sistem penjualan dan pengambilan keuntungan agar tidak mengalami kerugian terus-menerus.



- Heatmap profit berdasarkan state
  TagMedia menjual produknya ke beberapa *state* atau negara bagian yang beragam, berikut ini merupakan *heatmap* untuk menunjukkan *state* berdasarkan jumlah keuntungan dan penjualan. Pada grafik di bawah ini, dapat dilihat bahwa California dan New York memiliki jumlah penjualan yang paling besar, dibuktikan dari diagram dengan warna merah dan oranye. Apabila dibandingkan dengan jumlah *profit*, keuntungan yang didapatkan dari setiap *state* cenderung kecil dan bahkan mengalami ketugian di beberapa negara bagian seperti Texas danPensylvania. Analisis lebih lanjut dibutuhkan untuk mengetahui penyebab kerugian pada negara bagian tersebut. Perluasan pasar dibutuhkan untuk meningkatkan penjualan dan keuntungan pada negara bagian lainnya.

Modelling

Setelah melakukan persiapan data dalam bentuk eksplorasi, tahap selanjutnya data transaksi TagMedia dibuat ke dalam suatu model klasifikasi dan kluster. Berikut ini merupakan penjabaran dari setiap algoritma dan model yang dibentuk:

A. Model Klasifikasi

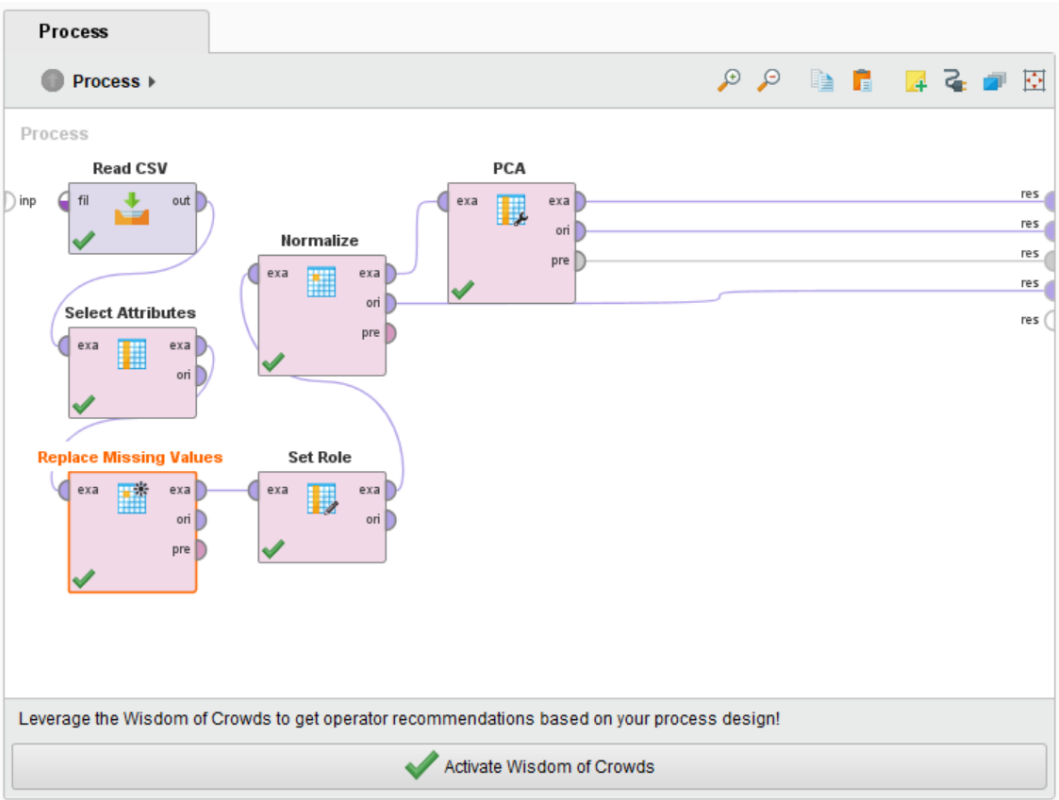
Klasifikasi merupakan salah satu teknik pembelajaran mesin terbimbing (*supervised learning)* yang digunakan untuk memprediksi kelompok dari suatu observasi baru berdasarkan data-data yang telah ada sebelumnya. Dalam teknik klasifikasi, atribut-atribut data dipecah menjadi *label* dan *feature*. *Feature* merupakan atribut deskriptif yang digunakan sebagai bahan prediksi, sedangkan *label* merupakan atribut yang berisi kelompok. *Label* ini yang akan dijadikan sebagai target atau hal yang ingin dicapai dalam proses klasifikasi, teknik klasifikasi akan memprediksi hasil *label* berdasarkan *input feature*.

Pada kasus data TagMedia, data terbaru memiliki 30 kolom seperti yang tertulis pada tabel di bawah. Data tersebut didapatkan dari transaksi dan sedikit penambahan *feature* (*feature engineering*) sebelum memasuki tahap pengolahan di RapidMiner. Tidak seluruh data dalam *data set* dapat digunakan untuk analisis, terdapat beberapa fitur yang kurang penting atau tidak mendukung hasil analisis. Atribut-atribut tersebut perlu diseleksi lewat tahap *feature selection* sebelum *modelling*.

Atribut					
OrderID	OrderDate	ShipDate	ShipMode	CustomerID	CustomerName
Gender	Age	ProductID	Category	SubCategory	ProductName
Quantity	PriceEach	Sales	Discount	TotalPrice	OriginPrice
Profit	BranchID	Segment	Country	City	State
PostalCode	Region	ShipDuration	OrderDay	Is_wekeend	Is_profitable

Pada kasus ini, klasifikasi dilakukan untuk keperluan segmentasi pelanggan. Segmentasi pelanggan merupakan teknik mengelompokkan pelanggan dalam kategori tertentu berdasarkan atribut lain yang tersedia dalam data. Sesuai dengan tujuan ini, *label* atau target pada klasifikasi ini adlaah variabel *Segment* yang terdiri atas tiga jenis pelanggan, yaitu *customer*, *home office*, dan *corporate*. Segmen pelanggan diketahui berdasarkan ciri-ciri yang dibawa oleh kelompoknya, proses segmentasi pelanggan ini dimulai dengan memilih dan menggunakan variabel-variabel yang bernilai penting atau berkontribusi besar. Sebelum memasuki tahap klasifikasi dengan algoritma *machine learning*, *feature selection* dilakukan dengan dua cara, yaitu *Principal Component Analysis* (PCA) dan Chi-Square.

- *Principal Component Analysis* (PCA)
  PCA merupakan metode yang digunakan untuk mengetahui fitur apa yang penting dalam sebuah *data set*. PCA umumnya digunakan untuk mereduksi dimensi dari *data set* yang besar sehingga menjadi kelompok data yang praktis untuk diimplementasi. Reduksi ini dilakukan sehingga pembuat model dapat meningkatkan akurasi model dengan menggunakan variabel-variabel terpilih dari *data set* yang berpengaruh secara signifikan. Secara khusus, RapidMiner memiliki operator PCA yang langsung dapat digunakan, berikut merupakan rancangan proses PCA dalam kasus TagMedia:



Tahap pertama dalam PCA adalah memasukkan data ke dalam *workspace* dengan operator *Read CSV*. Data yang sudah dimasukkan tersebut masih berupa data lengkap yang terdiri atas variabel numerik dan kategorikal. Proses PCA hanya dilakukan untuk variabel numerik saja, dengan demikian atribut numerik diseleksi dengan menggunakan operator *Select Attributes*. PCA tidak dapat dijalankan apabila ditemukan *missing value* dalam data, dengan demikian digunakan operator *Replace Missing Value* untuk mengubah nilai yang hilang menjadi nilai rata-rata.

Selanjutnya atribut *Set Role* dimasukkan ke dalam *workspace* untuk mendeklarasi variabel *Segment* sebagai sebuah *label* atau target. Selanjutnya dilakukan normalisasi terhadap data dan proses PCA dijalankan. Hasilnya adalah sebagai berikut:

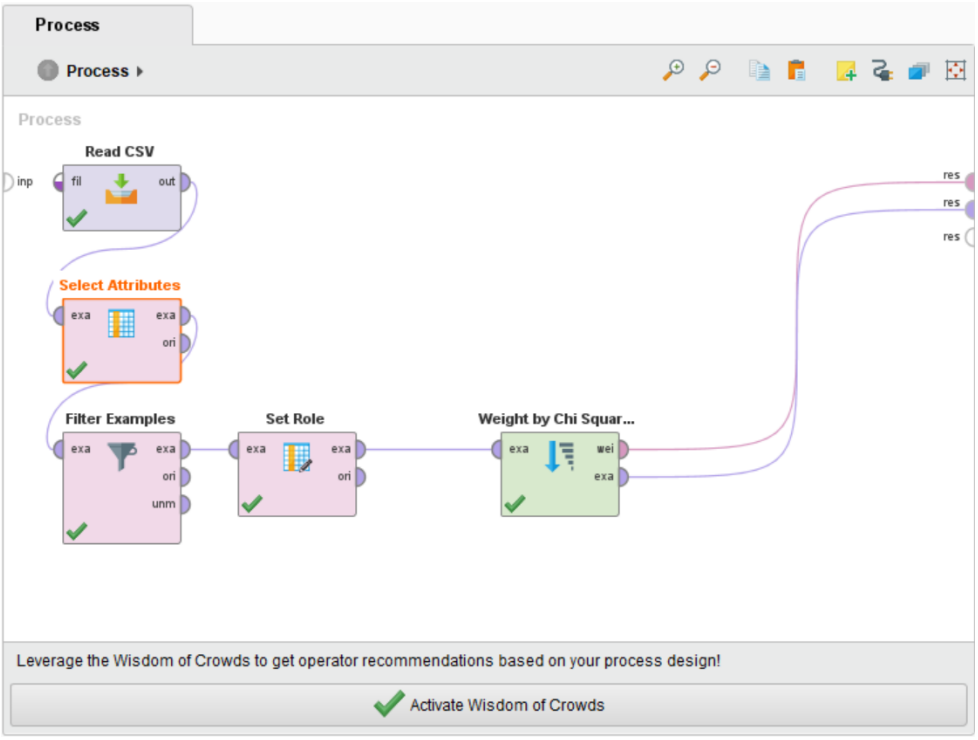
Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	2.045	0.465	0.465
PC 2	1.089	0.132	0.597
PC 3	1.039	0.120	0.717
PC 4	0.999	0.111	0.827
PC 5	0.910	0.092	0.919
PC 6	0.754	0.063	0.983
PC 7	0.381	0.016	0.999
PC 8	0.108	0.001	1.000
PC 9	0.000	0.000	1.000

PCA dijalankan dengan *variance threshold* sebesar 95%, hasil PCA yang dapat diambil adalah PC 1 hingga PC 5 karena mampu menjelaskan hampir 95% dari varians yang ada. Secara lebih jelas, berikut ini merupakan variabel yang terlibat dalam PCA. Variabel yang dapat digunakan untuk klasifikasi adalah *age, quantity, discount, sales, profit,* dan *ship duration.*

Attribute	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
Age	-0.280	0.101	0.316	-0.110	-0.587	0.668	-0.093	0.004	-0.000
Quantity	-0.083	-0.047	0.891	-0.169	0.206	-0.263	0.239	-0.001	0.000
Priceeach	-0.467	0.033	-0.187	0.044	-0.001	-0.035	0.536	0.675	0.000
Sales	-0.474	-0.002	0.011	0.012	0.126	-0.158	-0.482	0.049	0.707
Discount	0.026	0.754	0.031	0.032	0.544	0.364	-0.004	0.011	-0.000
Totalprice	-0.474	-0.003	0.011	0.012	0.126	-0.158	-0.482	0.049	-0.707
Originprice	-0.434	0.282	-0.154	0.037	-0.181	-0.247	0.388	-0.680	-0.000
Profit	-0.242	-0.582	-0.070	0.018	0.500	0.489	0.181	-0.277	-0.000
ShipDuration	0.007	-0.023	0.204	0.977	-0.054	0.023	0.001	-0.001	-0.000

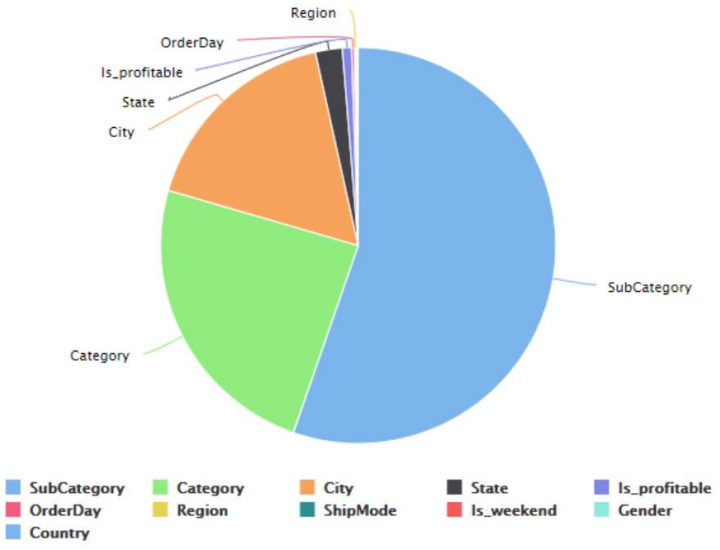
- Chi-Square

  Ch-Square merupakan salah satu cara untuk menunjukkan hubungan antara dua variabel kategorikal. Chi-Square dihitung dengan cara membandingkan perbedaan antara frekuensi yang diobservasi dengan frekuensi yang diharapkan untuk setiap atribut. Berikut merupakan desain proses untuk Chi-Square TagMedia:



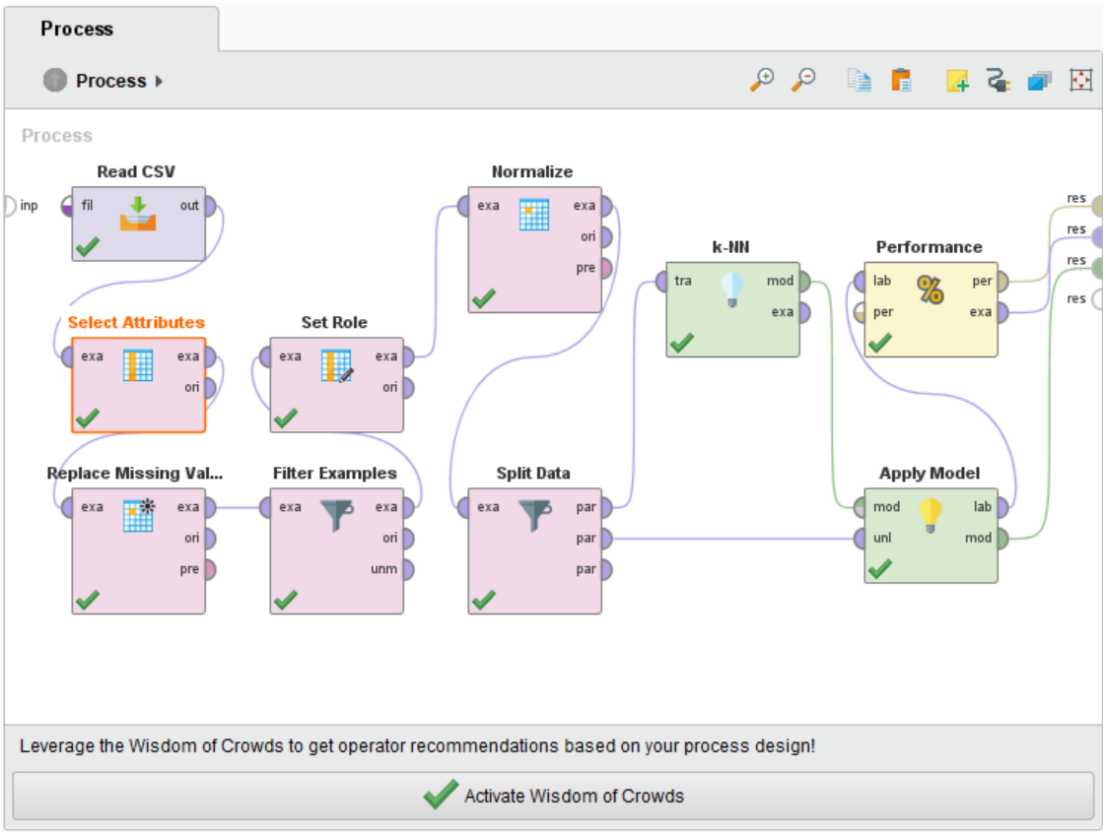
Setelah data dimasukkan ke dalam *environment*, data tersebut diseleksi kembali sehingga tidak memiliki *missing value* menggunakan *filter example*. Pada bagian *set role*, atribut *segment* dijadikan sebagai *label*. Hasilnya adalah sebagai berikut:

attribute	weight
Country	0
Gender	1.712
Is_week...	2.898
ShipMode	6.691
Region	11.083
OrderDay	16.478
Is_profit...	53.784
State	160.003
City	1240.570
Category	1760.299
SubCate...	4031.549



Berdasarkan hasil tersebut, tidak ada variabel kategorikal yang memiliki pengaruh signifikan terhadap segmentasi pelanggan. *Country* bernilai 0 karena hanya memiliki satu *value*, yaitu United States. Setelah melalui tahap seleksi fitur, didapatkan bahwa variabel yang memiliki pengaruh terhadap *segment* adalah *age, quantity, discount, sales, profit,* dan *ship duration.*

Pada tahap klasifikasi, variabel-variabel terpilih tersebut akan digunakan sebagai *feature* untuk menentukan pembagian kelompok pelanggan. Klasifikasi dilakukan menggunakan algoritma K-Nearest Neighbors (KNN). KNN merupakan algoritma *supervised learning* yang mengklasifikasikan data berdasarkan tetangga terdekat dan suara mayoritas. Algoritma ini bekerja dengan mempelajari data *training* untuk menemukan *label* dari objek baru. Berikut merupakan rancangan proses yang dilakukan dalam klasifikasi tipe pelanggan:



Tahap pertama dalam proses ini adalah memasukkan data ke dalam *workspace* RapidMiner menggunakan operator *Read CSV*. Data tersebut masih berbentuk data lengkap, sehingga dilakukan pemilihan atribut-atribut tertentu yang berperan penting dalam pembuatan model, yaitu *segment*, *age*, *quantity*, *discount*, *sales*, *profit*, dan *ship duration*. Apabila atribut numerik memiliki *missing value*, maka akan diganti menjadi nilai rata-rata, apabila atribut kategorikal memiliki *missing value*, maka baris yang memiliki nilai *null* akan dibuang menggunakan *filter example*. Proses ini dilakukan agar model tidak dibentuk dari baris dengan *missing value*. Selanjutnya, operator *set role* ditambahkan untuk mengubah atribut *segment* sebagai *label* untuk *training* dan *testing* model. *Set role* memberi informasi kepada algoritma bahwa klasifikasi akan dilakukan berdasarkan target *segment*, sedangkan sisanya menjadi *feature*. Tahapan selanjutnya yaitu normalisasi yang bersifat opsional, tahapan ini mengganti nilai numerik menjadi rentang 0 hingga 1. Algoritma ini mempelajari data *training* dan melakukan pengetesan dengan data *testing* untuk memprediksi hasil segmentasi pelanggan. Pemisahan data *train* dan *test* dilakukan menggunakan operator *split data* dengan pembagian 70% *training* dan 30% *testing*. Tahap selanjutnya adalah pembentukan model K-Nearest Neighbor dengan jumlah K sebesar 5 dan evaluasi dari hasil prediksi yang telah dilaksanakan.

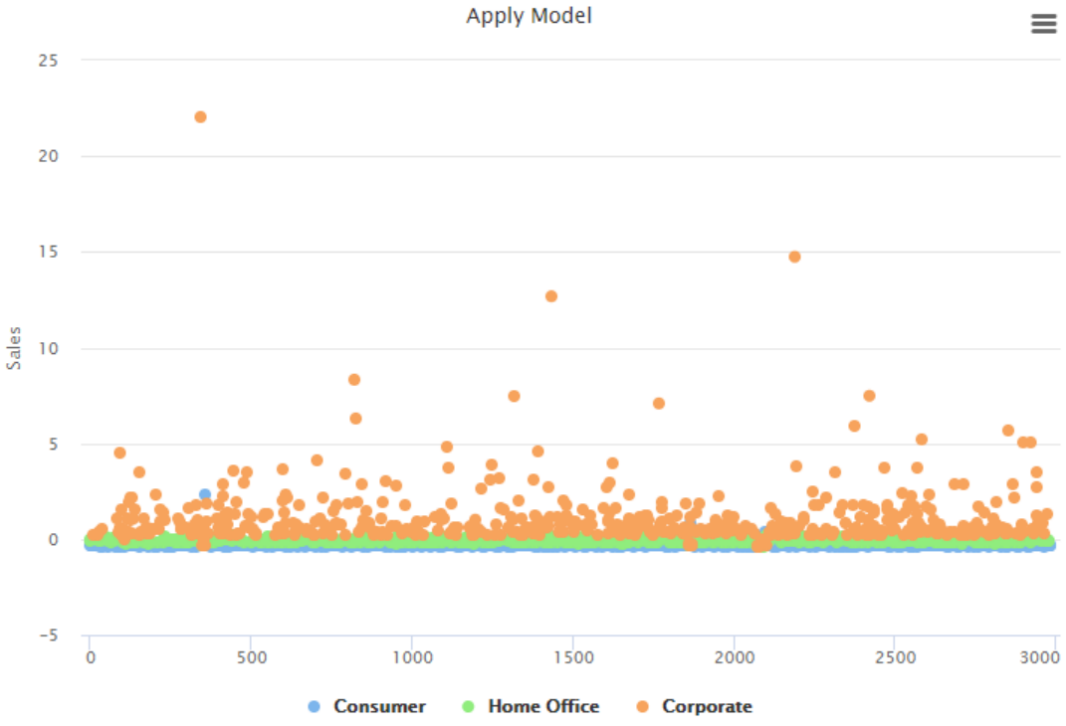
KNNClassification

```
Weighted 5-Nearest Neighbour model for classification.
The model contains 6966 examples with 5 dimensions of the following classes:
  Consumer
  Home Office
  Corporate
```

accuracy: 97.75%

	true Consumer	true Home Office	true Corporate	class precision
pred. Consumer	1842	12	0	99.35%
pred. Home Office	1	611	44	93.14%
pred. Corporate	0	10	464	97.89%
class recall	99.95%	96.52%	91.34%	

KNN memprediksi pembagian kelompok pelanggan berdasarkan *feature* dan melakukan evaluasi hasil dengan hasil di atas. Klasifikasi terbagi menjadi tiga jenis dengan akurasi prediksi sebesar 97,75%. Dalam model ini, masih terdapat beberapa kesalahan prediksi, namun secara umum memiliki tingkat akurasi yang sangat baik. Di bawah ini merupakan visualisasi klasifikasi tiga jenis pelanggan tersebut. Secara umum, model ini baik untuk digunakan dalam memprediksi segmentasi pelanggan. Umumnya, pelanggan dengan tipe *customer* melakukan pembelian dalam jumlah yang sedikit dan memberikan hasil *sales* yang lebih kecil dibandingkan *home office* dan *corporate*. Selain itu, pelanggan dengan tipe *customer* berada pada rentang usia yang lebih kecil dibandingkan tipe-tipe lain. Tipe *corporate* memberikan lebih banyak kontribusi terhadap jumlah *sales* yang besar dibandingkan dengan tipe lain.



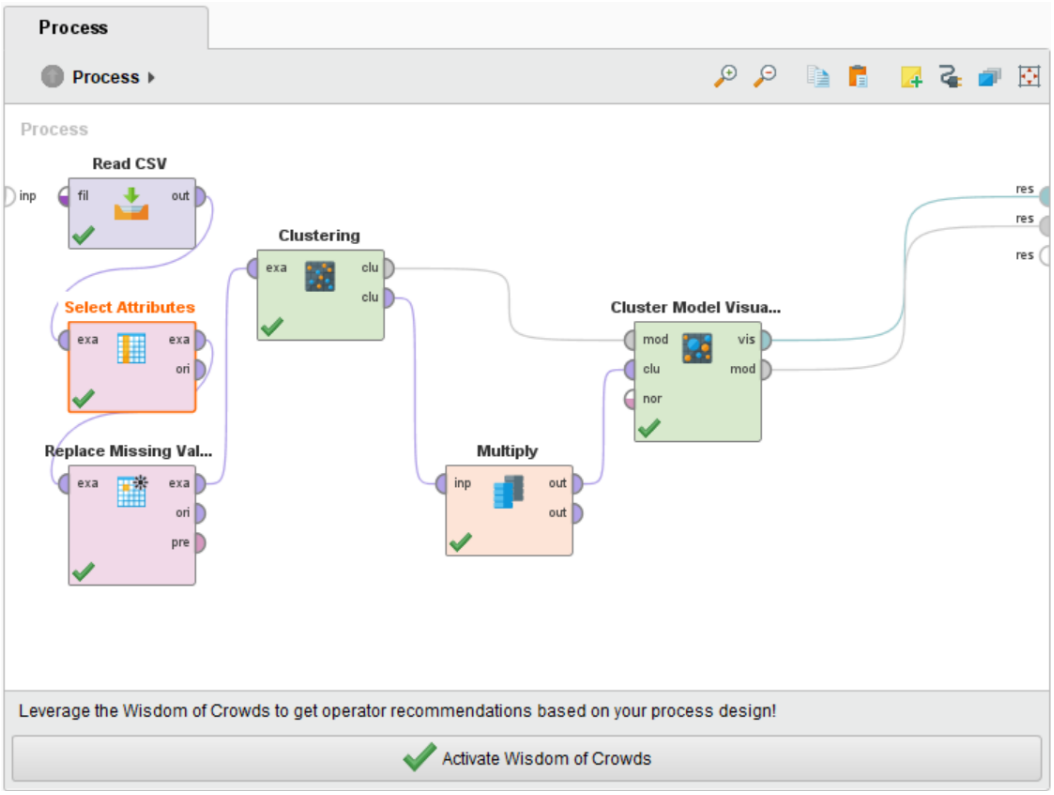
Pengenalan terhadap pelanggan merupakan salah satu langkah yang baik dalam menentukan strategi bisnis selanjutnya. Dengan mengetahui klasifikasi pelanggan ini, TagMedia dapat memberikan tawaran dan promosi yang berbeda untuk setiap tipe dengan ekspektasi *return value* yang berbeda-beda (*personalized*). Apabila TagMedia ingin mendapatkan jumlah penjualan yang besar dalam waktu singkat, disarankan untuk menargetkan penjualan kepada *corporate*. Jenis-jenis produk yang sering dibeli oleh *corporate* perlu dianalisis lebih lanjut untuk menawarkan barang yang serupa kepada pelanggan *corporate* baru. Selanjutnya apabila TagMedia memiliki data observasi baru, maka dapat menggunakan model ini untuk memprediksi kategorinya.

B. Model Kluster

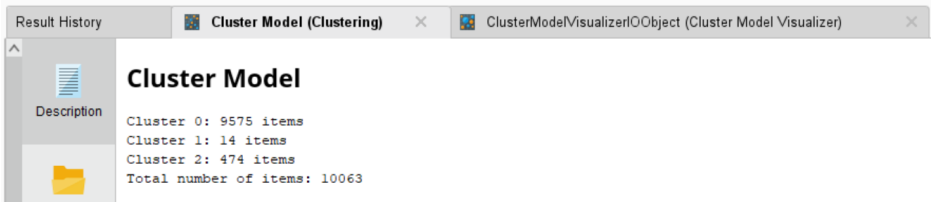
Clustering merupakan salah satu teknik *unsupervised learning* (pembelajaran mesin tidak terbimbing). *Clustering* digunakan untuk menemukan kelompok yang berharga di dalam sebuah data. Proses *clustering* berbeda dengan klasifikasi sebelumnya, *clustering* dibuat tanpa menggunakan *label* atau target apapun untuk dipelajari dari data *training*. *Clustering* membuat kelompok-kelompok data baru sesuai dengan kemiripan karakteristik data. Hasil akhir dari *clustering* adalah statistika deskriptif dari *cluster centroid* dan visualisasi grafik sesuai dengan pembagian kategori. Algoritma *clustering* yang digunakan untuk kasus TagMedia adalah K-Means. K-Means merupakan tipe kluster berbasis *prototype* dimana setiap data dipisah menjadi k kluster, dimana k merupakan jumlah *cluster* yang didefinisikan oleh pembuat model. Tujuan dari K-Means adalah untuk menemukan pusat data dan membagi data-data ke dalam kelompok yang berbeda. Secara teori, cara kerja dari K-Means adalah sebagai berikut:

1) Menentukan sentroid atau titik pusat data secara acak sebanyak k.
2) Setelah titik sentroid ditentukan, seluruh data dikelompokkan berdasarkan kedekatanya dengan sentroid tertentu. Kedekatan ini dihitung dengan jarak Euclidean. Tahap ini terjadi pemisahan data awal ke dalam kelompok sejumlah.
3) Setelah data berkumpul dan membentuk kelompok dengan sentroid sebelumnya, titik sentroid dihitung ulang untuk menemukan titik data yang baru. Setiap poin data akan kembali dipecah ke dalam kelompok-kelompok berdasarkan kedekatannya dengan titik sentroid.
4) Lakukan perulangan tahap 1-3 hingga seluruh kelompok terbentuk dengan sempurna.
5) Perulangan berhenti apabila tidak terdapat perubahan dalam pengelompokkan setiap poin data atau tidak ada perubahan signifikan pada titik sentroid.

Tidak seperti *classification*, *cluster* tidak memerlukan *label* atau pemisahan data *training-testing* sehingga kerangka proses dalam RapidMiner akan terlihat lebih sederhana seperti di bawah ini:



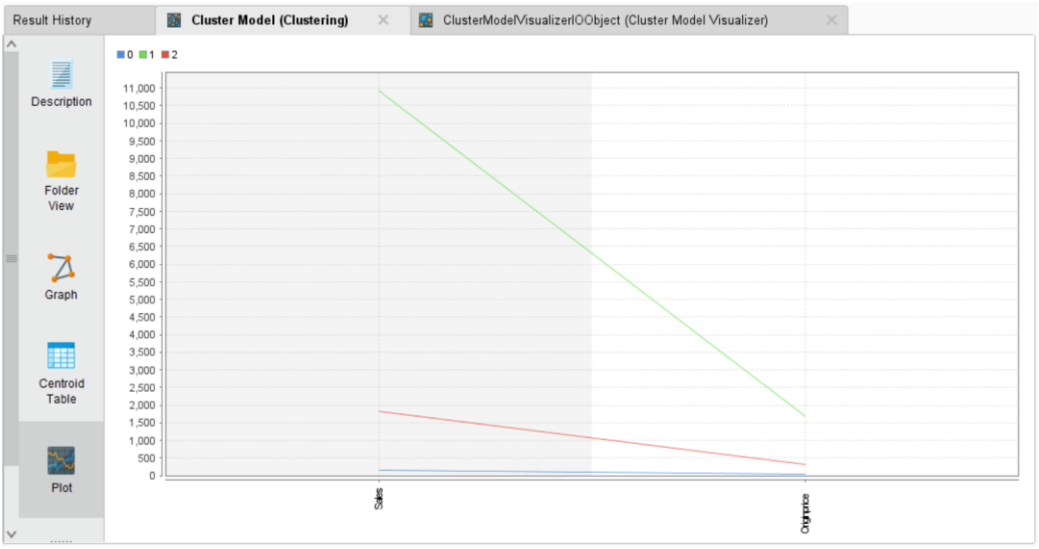
Pertama-tama, data TagMedia dimasukkan ke dalam *workspace*, kemudian dipiih dua atribut numerik yang ingin dijadikan dasar pengelompokkan. Pada kasus ini, TagMedia ingin mengelompokkan harga beli produk dengan nilai *sales* (penjualan produk). TagMedia ingin melihat kategori harga rendah, sedang, dan tinggi. Tahap selanjutnya adalah mengganti *missing value* pada atribut numerik apabila ada. Setelah itu, *clustering* dilakukan dengan algoritma K-Means dengan jumlah k = 3 dan maksimal *run* = 20. Hasil dari kluster tersebut ditampilkan melalui *Cluster Model Visualization*. Hasilnya adalah sebagai berikut, *cluster 0* merupakan kategori rendah, *cluster 1* merupakan kategori tinggi, dan *cluster 2* merupakan kategori sedang.



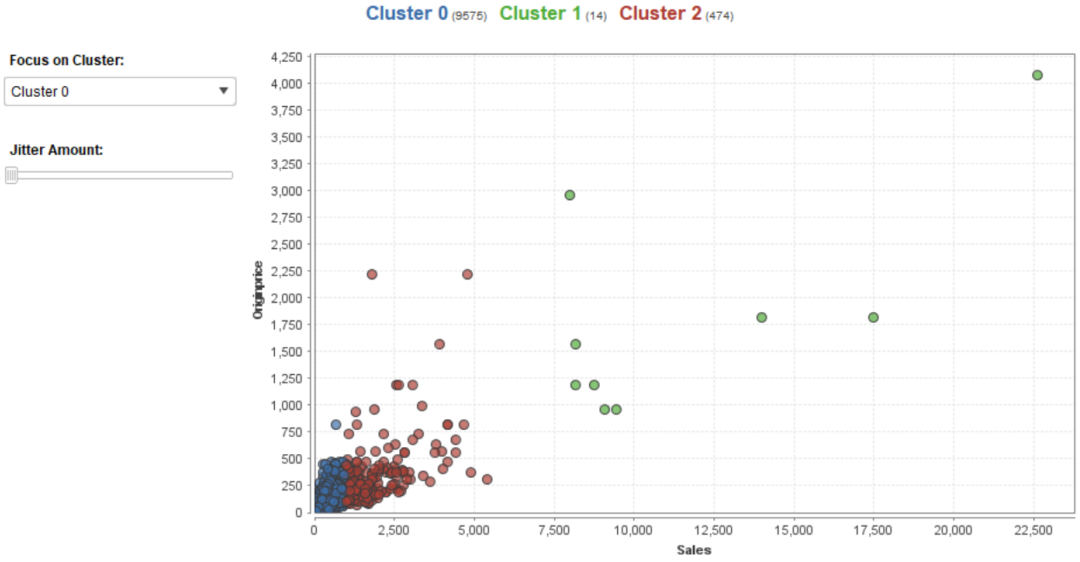
Berikut ini merupakan *centroid table* untuk masing-masing nilai titik sentroid setiap kelompok.

Cluster	Sales	Originprice
Cluster 0	137.632	38.004
Cluster 1	10912.744	1684.517
Cluster 2	1816.156	318.660

Berikut ini merupakan *centroid chart* setiap kelompok.

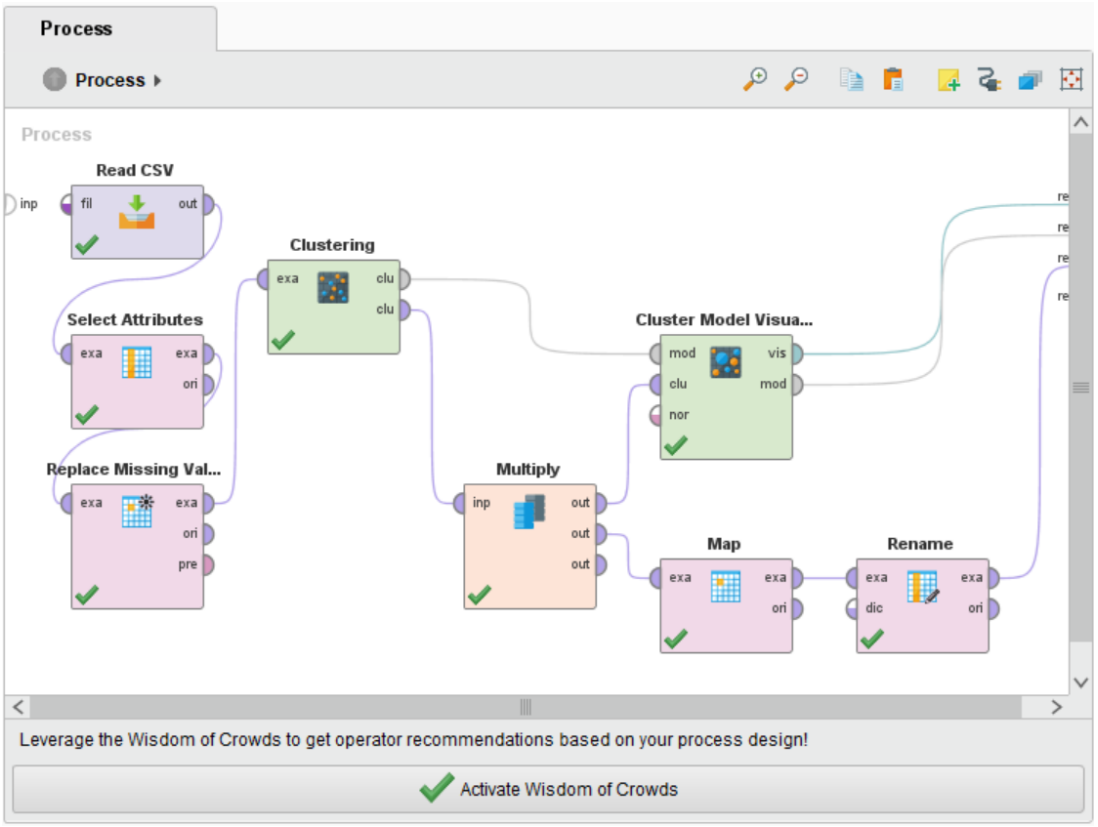


Melalui diagram di bawah ini, diketahui bahwa semakin besar harga beli produk dari pemasok, maka nilai jual juga akan semakin tinggi. Pada kelompok rendah, dapat dilihat bahwa beberapa produk dijual lebih rendah dibandingkan harga belinya, hal menjadikan TagMedia menjadi kehilangan keuntungan. Transaksi pada kelompok ini perlu dievaluasi sehingga tidak mengalami kerugian berkepanjangam. Pada kategori sedang, produk-produk yang memiliki harga di bawah $2000 dapat dijual dengan harga di atas $2000. Data poin untuk data tersebut tergolong banyak, artinya banyak pelanggan yang bersedia untuk membayar produk dalam rentang harga tersebut. Pembeli dalam kelompok ini perlu dianalisis lebih lanjut dan ditawarkan produk-produk dengan harga yang serupa, TagMedia juga dapat meneruskan pengambilan barang dan harga jual sesuai dengan kelompok ini. Pada *cluster* kategori tinggi, pengambilan keuntungan mencapai 4-5 kali lipat dari harga beli. Produk dijual dalam harga yang jauh lebih tinggi sehingga menghasilkan *profit* yang juga sangat tinggi bagi TagMedia, namun kelompok ini hanya memiliki sedikit transaksi.



Secara umum, produk terbagi menjadi tiga kelas yang berbeda, yaitu rendah, sedang, dan tinggi yang berguna untuk mencari tahu kemungkinan *profit* yang dapat diraih berdasarkan nilai beli dan nilai jual produk. Pembagian ini juga dapat digunakan sebagai pendukung dalam pengambilan keputusan untuk menetapkan atau menaikkan

harga jual produk, serta mengetahui pelanggan mana yang cocok untuk dijadikan target dengan *profit* tinggi. Berikut ini merupakan proses tambahan (opsional) untuk mengganti *value* cluster_0 menjadi "rendah", cluster_1 menjadi "tinggi", dan cluster_2 menjadi "sedang" menggunakan operator *Map*. Selain itu, nama kolom hasil prediksi juga diubah menggunakan operator *Rename*.



Hasil dari proses di atas adalah sebagai berikut, kolom SalesCat dapat digabungkan ke dalam *data set* asli sehingga tiap transaksi memiliki label tambahan, yaitu kategori penjualan. Melalui pembagian ini dapat dianalisis lebih lanjut mengenai pelanggan yang memberikan nilai *sales* tinggi dan tipe barang yang memiliki *origin* price tinggi.

Row No.	id	SalesCat	Sales	Originprice
1	1	rendah	75.880	20.110
2	2	rendah	1.248	1.060
3	3	rendah	78.304	24.470
4	4	rendah	21.456	1.610
5	5	rendah	157.920	28.030
6	6	rendah	203.184	93.970
7	7	rendah	58.380	4.590
8	8	rendah	105.520	14.250
9	9	rendah	80.880	9.980
10	10	rendah	5.980	3.290
11	11	rendah	899.136	196.690
12	12	rendah	71.760	8.610

ExampleSet (10,063 examples, 2 special attributes, 2 regular attributes)

Referensi:

Dahman, M. (2018). Univariate Descriptive Statistics-Chapter Two. 10.31219/osf.io/thd3c

Materi Perkuliahan *Big Data Analytics 2* Pertemuan 2-6

Spolaor, N., Cherman, E., & Monard, M. C., & Lee, H. (2013). A Comparison of Multi-label Feature Selection Methods using the Problem Transformation Approach. *Electronic Notes in Theoretical Computer Science*, 292, 135–151. 10.1016/j.entcs.2013.02.010.