04

Medidas

Medidas

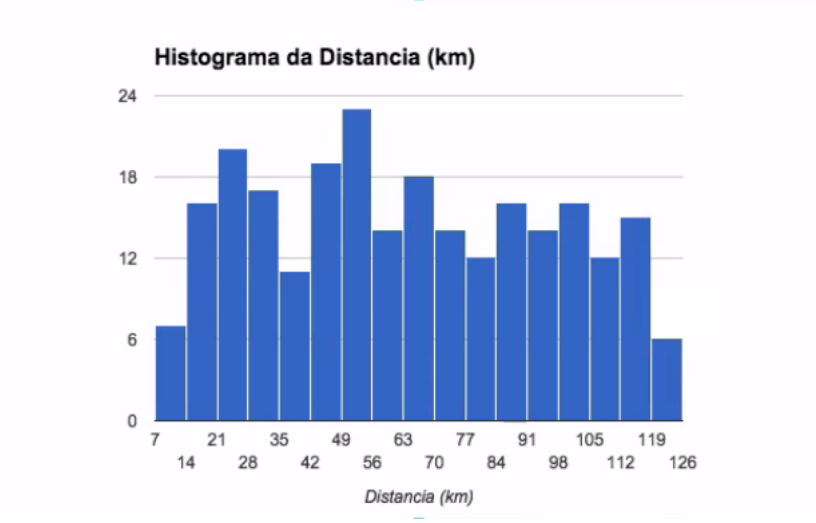
Continuando o trabalho de análise, você recebeu um novo desafio: Verificar os dados sobre o tempo do serviço de entrega da Jump Cats. Você precisará descobrir se o tempo é regular, se a demora é grande, qual é a velocidade média. A empresa quer ter uma ideia de como o serviço está se comportando.

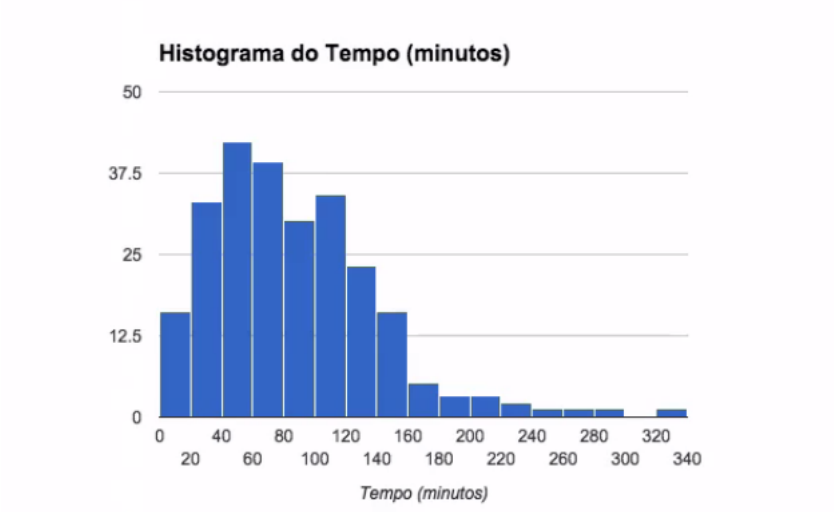
Foi fornecida uma planilha com a distância e o tempo.Nossa tarefa é entender para chegar a algum valor ou estatísticas úteis para a empresa.

Distancia (km)	Tempo (minutos)
29	78
27	32
105	120
70	144
120	116
85	119
38	43
26	36
24	31
21	38
33	39
32	36
52	79
36	56
53	56
18	20
13	15
68	98
77	71

A planilha possui a distância em quilômetros e o tempo em minutos da entrega. Durante a semana, eles coletaram informações 250 eventos. Primeiramente, isto não é uma série temporal. Não estamos analisando por datas, os valores estão soltos. A maneira como vamos tratar os dados será diferente da maneira como trataríamos uma série temporal. Poderíamos tratá-los como uma série temporal, se tivéssemos disponível o dia e a hora? Sim. Mas seriam informações irrelevantes para nós.

Iremos aqui trabalhar apenas como um conjunto de valores. Geralmente, nestes casos a primeira coisa que observamos é a distribuição. Ou seja, como os valores estão distribuídos, quais são os mais e menos comuns. Existe um gráfico chamado **histograma**, que irá desenhar justamente isto. Criaremos um para `Distância` e outro para o `Tempo`.





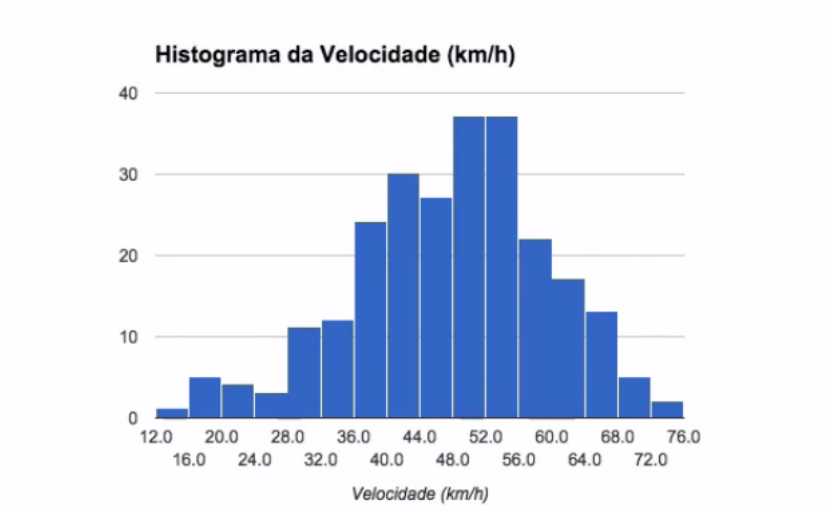
O eixo `Y` sempre será referente a contagem.

Observe que as barras da distância são uniformes. No entanto, o histograma do tempo tem uma maior concentração até o tempo de 160 minutos e depois uma "cauda". Como analista nós temos que investigar o motivo de termos entregas que demoram até 340 minutos. Os motivos podem ser variados: a moto quebrou, o trânsito estava intenso ou tinha uma greve nas vias. É muito provável que algo aconteceu nestes eventos. Será preciso a coleta de novos dados para responder esta questão.

Nós também poderíamos calcular a velocidade média das motos durante a entrega. Encontramos estes valores dividindo a `Distância` pelo `Tempo`. Nossa sugestão é que você converta o tempo de **minuto** para **hora**.

Velocidade (km/h)
22.2
50.0
52.5
29.3
62.1
42.9
53.4
43.1
47.1
33.5
50.6
53.6
39.3
38.4
56.8
53.2
52.4
41.7
65.0

Criaremos também o histograma da `Velocidade` .



Agora, vemos uma cauda no lado esquerdo do histograma. Isto significa que alguns entregadores foram a uma velocidade inferior a 30km/h. Provavelmente existiu algum problema, porque não é normal um entregador ir a 12km/h, por exemplo. Analisando também o gráfico da distância, vemos que tiveram entregas feitas em 120 km. Se a entrega foi feita em uma estrada, é normal que a velocidade seja superior a 60km/h.

As primeiras análises que podemos fazer é: a distância é constante, a maior parte das entregas são feitas em até 160 minutos, o que equivale a duas horas e meia. Vemos também que algumas entregas demoraram muito para serem feitas. Também percebemos que algumas entregas tiveram uma velocidade média muito baixa, enquanto a maior parte das entregas foram feitas com velocidades entre 40 e 68km/h.

Continuando a análise dos dados

Existe uma maneira muito usada para análise, além dos histogramas, e que acaba resumindo os dados. Existirão casos nos quais teremos que trabalhar com mais de 10 mil dados e não fará sentindo lermos todos eles. Precisamos resumi-los em poucos valores, que ajudem a explicar determinados comportamentos. Números que serão usados como guias para se ter uma visão rápido do que está acontecendo. Basicamente, eles terão os seguintes objetivos.

Valor Mínimo
1o Quartil
2o Quartil
3o Quartil
Valor Máximo
Média
Moda
Variância
Desvio Padrão

Nós iremos começar com um pequeno conjunto de dados que contem 15 informações, ordenadas em ordem crescente.

1	2
2	8
3	12
4	25
5	31
6	31
7	41
8	46
9	49
10	56
11	63
12	70
13	71
14	84
15	95

Iremos preencher os nossos objetivos com os dados. É fácil identificar quais são os valores mínimos e máximos. Nós iremos explicar o que aconteceu no "meio" da lista de dados, com três números que chamaremos de **quartil**, que é referente a quarta fração. Para encontrá-los, iremos dividir os números em quatro blocos. De `1` até `15`, o número posicionado no meio é `48` e será o `2o quartil`. O `1º quartil` será o que está no primeiro quarto da lista, assim como o `3º quartil` será o que estará no terceiro quarto.

1	2	
2	8	
3	12	
4	25	1 Quartil
5	31	
6	31	
7	41	
8	46	2 Quartil
9	49	
10	56	
11	63	
12	70	3 Quartil
13	71	
14	84	
15	95	

Já conseguimos preencher o resumo com alguns dados.

Valor Mínimo	2
1o Quartil	25
2o Quartil	46
3o Quartil	70
Valor Máximo	95
Média	
Moda	
Variância	
Desvio Padrão	

Percebemos que 25% dos dados estão abaixo de `25`, metade estão abaixo de `46`, 75% dos dados estão abaixo de `70`. E o maior valor é `95`. O valor mínimo também recebe o nome de `0º quartil` e o máximo também é chamado de `4º quartil`. O valor que está no meio também recebe o nome de `mediana`.

Iremos preencher os valores que faltaram no resumo: a `média` se refere à média aritmética. Neste caso ela irá coincidir com a mediana. O conceito de `moda` trata dos valores que aparecem mais vezes listados, no nosso caso, será `31`. Na prática, a `moda` é pouco utilizada.

O números chamados de `Média`, `Mediana` e `Moda` são chamados **números de tendência central**. Eles indicam que os números da lista estão distribuídos entorno de um ponto. A `variância` padrão mostrará o quão dispersos os números estão. Para calculá-la iremos usar a função `VAR` do Spreedsheet e selecionar os dados da lista. O resultado será `778,114`. Também podemos calcular o desvio padrão usando o `STD` (*Standard deviation*, em inglês) ou encontrando a raiz quadrada da variância, usando a função `SQRT` (*square root*).

No fim, ficamos com os dados preenchidos da seguinte forma:

Valor Mínimo	2
1o Quartil	25
2o Quartil	46
3o Quartil	70
Valor Máximo	95
Média	46
Moda	31
Variância	778.1142857
Desvio Padrão	27.89469996

Ele servem para descrever os valores com que trabalhamos.

1	2	0 Quaril	Valor Mínimo
2	8		
3	12		
4	25	1 Quartil	
5	31		
6	31		
7	41		
8	46	2 Quartil	Mediana
9	49		
10	56		
11	63		
12	70	3 Quartil	
13	71		
14	84		
15	95	4 Quartil	Valor Máximo

Do resumo, geralmente é levado tão em consideração a `moda` e a `variância`. Já o `desvio padrão`, sim, é relevante. No caso, encontramos o valor `27,89`, o que significa que a dispersão do centro tem este valor. Este dado pode ser usado para compararmos esta distribuição com outra.

Resumindo os números da entrega

Agora que usamos um exemplo simples para compreender os conceitos, vamos para o caso dos números das entregas. Temos uma tabela com informações semelhantes as que levantamos anteriormente. Vamos preenche-las com os números das colunas de `Distancia`, `Tempo` e `Velocidade`. Para o valor mínimo, usaremos a função `Min`, para os quartils, usaremos `QUARTILE` e especificaremos a qual nos referimos (primeiro, segundo ou terceiro). O valor máximo encontraremos com a função `MAX`, a média, com a `AVERAGE`, e para o desvio padrão, utilizaremos `STDEV`.

	Distancia (km)	Tempo (minutos)	Velocidade (km/h)
Valor Mínimo	10	8.8	14.8
1o Quartil	35	47.5	40
2o Quartil/Mediana	61	77.5	49
3o Quartil	92	117.2	55
Valor Máximo	120	336.2	76.0
Média	63.48	86.2	47.8
Desvio Padrão	31.6	53.1	11.6

Analisando os valores, veremos que a média e mediana da `Distancia` estão próximas. Isto significa que o histograma está bem distribuído, sem deslocamentos para esquerda ou direita. Vemos que 25% das pessoas estão a uma distância entre 92 e 120 km. Talvez, seja o caso de criar um ponto de distribuição em outro local. Também, temos uma quantidade de 25% de entregas com um distância inferior a 35 km.

Passando para a coluna `Tempo` , percebemos que o valor mínimo é 8 minutos, mas o 1º quartil foi para 47 minutos. Até o 3º quartil, ou seja 75% dos casos, já chegaremos a 117 minutos. O histograma será deslocado para esquerda. A diferença do valor do 3º quartil para o valor máximo é superior ao dobro. Também é possível notar que 50% das entregas são feitas entre 1h e 2h. Significa que em uma jornada de 8 horas, um entregador fará de 4 a 8 entregas por dia. Observando o histograma, é preferível dizer que a média de tempo está mais próxima do valor da mediana, porque o gráfico tende para a esquerda. Mas existem alguns casos extremos, nos quais a entrega demorou muito tempo.

Ao observarmos os dados da terceira coluna, referente à `Velocidade` . Vemos que o valor mínimo foi 14,8 km/h. Também percebemos que o 1º, o 2º, e o 3º quartil estão próximos, com apenas 15 km/h de variação. Percebemos esta concentração dos dados no histograma. O valor máximo é 76 km/h.

O **desvio padrão** nos dá um ideia de espalhamento das informações. Quando ele será útil? Quando repetirmos o experimento e tivermos uma tabela semelhante, com novos valores de desvio padrão. Dependendo da distribuição, pode resultar em alguns percentuais. Como estamos analisando os dados empiricamente, o desvio é usado para comparação. Se repetíssemos a experiência, observaríamos se houve ou não um aumento no desvio padrão. No caso de um aumento, significa que as os valores dos dados estão mais dispersos. Se houve um diminuição, significa que os valores estão mais concentrados e há mais uniformidade nos dados.

Com esta tabela, cumprimos a missão de resumir os dados. Podemos incluir estas informações da tabela em um relatório, e conseguimos representar de uma forma numérica o que foi apresentado nos histogramas.