

Ficha Técnica do Projeto 2 de análise de dados

Título Projeto: Análises de Hipóteses Spotify

Objetivo: Identificar padrões ou características que influenciam a popularidade de uma música, utilizando dados do Spotify.

Equipe: Projeto em duplas

Ester Linderos Saravia e Daiane Sinzato.

Cronograma:

- Semana 1:
 - Marco 1:
 - Limpar os dados e unir as tabelas no BigQuery.
 - Gerar tabelas, gráficos, cálculos de quartis e correlação durante a análise exploratória.
 - Aprender Power Bi e exportar dados.
 - Aplicar a técnica de segmentação e validação de hipóteses.
 - Preparação para Marco 2:
 - Preparar ambiente no Google Colab.
 - Criar conta e preparar repositório no GitHub.
- Semana 2:
 - Marco 2:
 - Importar os dados para o Google Colab.
 - Realizar testes de normalidade (Shapiro Wilk)
 - Realizar testes de significância (teste T ou Wilcoxon).
 - Interpretar os resultados e criar uma tabela resumo.
 - Integração com Marco 1:
 - Adicionar os resultados à apresentação do Marco 1.
- Semana 3:
 - Marco 3:
 - Realizar análise de regressão linear.
 - Interpretar os resultados e criar um gráfico.
 - Finalização do Projeto:
 - Adicionar os resultados à apresentação final.
 - Preparar e gravar vídeo explicativo.
 - Criar Ficha Técnica completa.
 - exportar google colab ao repositório no GitHub.

Recursos:

- BigQuery
- Google Colab
- Power BI
- Apresentações Google(use looker)
- Loom (para gravação de vídeo)
- GitHub (para armazenamento do projeto)

Observações: O projeto será desenvolvido ao longo de 3 semanas, com foco inicial na limpeza e análise exploratória dos dados, seguido pela realização de testes de significância e, por fim, pela aplicação da regressão linear. A apresentação final será feita por meio de um vídeo explicativo, onde serão apresentadas as conclusões e recomendações obtidas durante a análise dos dados do Spotify.

Links de Interesse:

Dataset: Link([Hipotese](#))

Power BI: Link Exploratoria([Hipoteses](#))

Power Bi: Link dashboard([Hipótese](#))

Apresentação: Link([Hipótese](#))

Loom Video: Link([Hipótese](#))

hipóteses levantadas:

- Músicas com BPM (Batidas Por Minuto) mais altas fazem mais sucesso em termos de número de streams no Spotify.
- As músicas mais populares no ranking do Spotify também possuem um comportamento semelhante em outras plataformas, como a Deezer.
- A presença de uma música em um maior número de playlists está correlacionada com um maior número de streams.
- Artistas com um maior número de músicas no Spotify têm mais streams.
- As características da música influenciam o sucesso em termos de número de streams no Spotify.

Para o desenvolvimento do projeto, seguir os passos:

Coleta e Preparação dos Dados: Utilize os dados do Spotify fornecidos pela gravadora. Faça a limpeza dos dados, tratando valores nulos, duplicados e outliers, e prepare os dados para análise no Bigquery.

- deixe fora as colunas mode, key e shazam charts por ficar fora do escopo e ter nulos.

Análise Exploratória dos Dados: Explore os dados para entender suas características. Analise a distribuição das variáveis(histograma), calcule medidas de tendência central(média, mediana) e dispersão(desvio padrão), e visualize os dados para identificar padrões e insights, no Power BI.

- identifique e trate valores discrepantes de variável string e numéricas(deixe sem símbolos no string e apaguei o streams nao numerico, trocando de object a int64), agrupar variáveis categóricas(total de playlists e data com ano-mês-dia de lançamento), crie 3 tabela limpa e junte elas.
- Categorizar variável(percentil: low, medium e high)

Validação das Hipóteses: Para cada uma das hipóteses levantadas, realize análises estatísticas adequadas para determinar se elas são válidas. Por exemplo, para a hipótese sobre BPM e número de streams, calcular a correlação entre essas variáveis.

- Medidas de tendência central e desvio padrão para variáveis numéricas e teste de correlação de Pearson.
- tabela de resumo(matriz) e gráfico de linha para visualizar dados ao longo do tempo.
- use gráfico de barra e de dispersão para visualizar hipótese.

Criação de Dashboard: Utilize o Power BI para criar um dashboard que apresente os principais insights encontrados durante a análise exploratória e a validação das hipóteses. O dashboard deve ser intuitivo e fácil de entender.

Apresentação dos Resultados: Prepare uma apresentação no Looker Studio que resuma os principais resultados da análise. Inclua gráficos, tabelas e insights importantes para comunicar suas descobertas de forma clara e convincente.

Testes de Significância: Se necessário, realize testes de significância estatística para validar suas conclusões e garantir a robustez dos resultados.

- realize o teste de shapiro wilk para ver se os dados têm uma distribuição normal
- Para confirmar ou refutar as hipóteses use o teste de Mann-Whitney U (porque os dados não tem uma distribuição normal) para a comparação entre grupos e de correlação de Pearson, para análise de correlação. Consideramos um nível de significância de 5% ($\alpha=0.05$) para todos os testes.

Regressão Linear: Se desejar explorar mais a relação entre variáveis, aplique a técnica de regressão linear para identificar a influência das características da música no número de streams.

Finalização e Entrega: Compile todos os resultados, análises e recomendações em um relatório final. Grave um vídeo de apresentação explicando suas conclusões e recomendações, utilizando o Loom, e compartilhe o vídeo e o relatório.