Relatório do Trabalho:

Definição de Dataset, Limpeza e Análise dos Dados

Melchiades Blanco Junior

Resumo

Relatório referente ao trabalho de limpeza e análise de dados colhidos do portal de dados abertos da prefeitura de Curitiba.

Relatório do Trabalho:

Definição de Dataset, Limpeza e Análise dos Dados

Este documento tem o objetivo de descrever as atividades executadas no jupyter notebook referente aos seguintes tópicos:

- Descrição do dataset (descreva o objetivo do dataset e os dados contidos nele como colunas, cobertura de dados, etc.)
- Procedimentos de Limpeza (descreva como os dados foram limpos, o que foi retirado e quantos dados sobraram no final)
- Análise Exploratória (apresente os gráficos e análises estatísticas elaborados; para cada elemento incluído, descreva quais padrões podem ser observados e sugira hipóteses para explicá-los)
- Perguntas iniciais (descreva as perguntas que você pretende responder nas análises sobre os dados. estas perguntas podem ser alteradas nas próximas etapas).

Descrição do Dataset

O dataset escolhido para este trabalho é o que contém as informações sobre as compras e contratações, com o detalhamento dos gastos públicos.

Órgão responsável: Administração e Gestão de Pessoal

Responsável:

Contratações: Patricia Mendes Maurer (Departamento de Gestão de Serviços)

Licitações: Fabíola Roberti Coneglian (Departamento de Licitação e Gestão de Compras)

Frequência de atualização: Mensal

Espectro temporal: Últimos 5 anos

Campos: [Ver na seção dicionário de dados]

Ultima atualização: 01/04/2021 09:12

Dataset encoding: ANSI "ISO-8859-1"

Separador de coluna: ';'

Separador decimal: ','

Dicionário de dados

Dataset Empenhos.

Este dataset possui informações como os valores empenhados e pagos para as licitações.

Campo	Descrição	Tipo de Dado	Tamanho
Órgão	Secretaria ou ente responsável pela realização da despesa	varchar	10
Número do Processo	Numeração sequencial do procedimento licitatório.	varchar	20
Modalidade	Forma de Contratação	varchar	50

Objeto	São bens e serviços cujos padrões de desempenho e qualidade são objetivamente definidos na instrução processual.	text	16
Valor Total/Global	Valor que integraliza os valores de todos os itens unitários contidos no processo de aquisição/contratação.	decimal	9
Local de Execução/Entrega	Endereço onde será entregue o objeto licitado.	varchar	70
Data Diário Oficial	Data de publicação dos atos no Diário Oficial do Município.	datetime	8
Protocolo	Conjunto de documentos oficialmente reunidos em uma única arquivística.	varchar	524
Início da Vigência do Contrato	Data inicial de vigência do contrato	datetime	8
Fim da Vigência do Contrato	Data final de vigência do contrato	datetime	8
Situação	Situação do andamento em que se encontra o contrato ou o empenho	varchar	50

Dataset Dados Básicos.

Este dataset possui informações básias sobre as licitações, por exemplo, data da licitação e orgão responsável.

Campo	Descrição	Tipo de Dado	Tamanho
Órgão	Secretaria ou ente responsável pela realização da despesa	varchar	10
Número do Processo	Numeração sequencial do procedimento licitatório.	int	4
Modalidade	Forma de Contratação	varchar	50
Número do Contrato	Número sequencial de todo e qualquer ajuste entre órgãos ou entidades da Administração Pública e particulares, em que haja um acordo de vontades para a formação de vínculo e a estipulação de obrigações recíprocas, seja qual for a denominação utilizada.	int	4

Número Empenho	Número sequencial de documento utilizado para registrar as despesas orçamentárias de atos emanados pela autoridade competente.	int	4
Contratado/Fornecedor	Pessoa física ou jurídica signatária de instrumento contratual com a Administração Pública, na condição de fornecedor de bens, executor de obra ou prestador de serviço.	varchar	70
CNPJ/CPF	Número sequencial do fornecedor/contratado, registrado junto à Receita Federal do Brasil.	varchar	18
Valor Empenhado	Consiste na reserva de dotação orçamentária para um fim específico.	decimal	9
Valor Liquidado	Valor verificado do direito adquirido pelo contratado.	decimal	17
Valor Anulado	São despesas empenhadas de exercícios financeiros ou do atual, que não serão mais pagas ao contratado.	decimal	17
Valor Pago	Valor efetivamente recebido pelo contratado.	decimal	17
Valor a pagar	São as despesas empenhadas, mas não pagas dentro do exercício financeiro.	decimal	17
Fonte de Recursos	As fontes de recursos constituem-se de determinados agrupamentos de naturezas de receitas, atendendo a uma determinada regra de destinação legal, indicando como são financiadas as despesas orçamentárias.	varchar	100

Dataset Itens do processo.

Este dataset possui os itens inclusos em cada licitação descritos de forma unitária.

Campo	Descrição	Tipo de Dado	Tamanho
Órgão	Secretaria ou ente responsável pela realização da despesa	varchar	10
Número do Processo	Numeração sequencial do procedimento licitatório.	int	4
Modalidade	Forma de Contratação	varchar	50
Item	Descrição do objeto contratado	varchar	121

Quantidade	Relação numérica que expressa a quantia necessária do objeto licitado.	decimal	9
Unidade de Medida	Unidade de medida do objeto contratado, conforme grandeza que compõe o sistema métrico decimal	varchar	2
Contratado/Fornecedor	Pessoa física ou jurídica signatária de instrumento contratual com a Administração Pública, na condição de fornecedor de bens, executor de obra ou prestador de serviço.	varchar	70
CNPJ/CPF	Número sequencial do fornecedor/contratado, registrado junto à Receita Federal do Brasil.	varchar	18
Número do Contrato	Número sequencial de todo e qualquer ajuste entre órgãos ou entidades da Administração Pública e particulares, em que haja um acordo de vontades para a formação de vínculo e a estipulação de obrigações recíprocas, seja qual for a denominação utilizada.	int	4
Início da Vigência do Contrato	Data inicial de vigência do contrato	datetime	8
Fim da Vigência do Contrato	Data final de vigência do contrato	datetime	8
Valor Unitário	Valor da unidade de medida.	decimal	9
Valor Total/Global	Valor que integraliza os valores de todos os itens unitários contidos no processo de aquisição/contratação.	decimal	9

Procedimentos de Limpeza

Durante o procedimento de limpeza, as etapas abaixo foram avaliadas e realizadas conforme foi necessário:

- Remoção de registros com valores NaN
- Remoção de registros com valores com valor inválido
- Conversão dos tipos das colunas
- Junção dos datasets

- Seleção das colunas que serão utilizadas
- Conversão de valores
- Remoção dos outliers
- Salvar arquivo limpo

Remoção de registros com valores NaN

Ao analisar os dataset, encontramos algumas colunas com valores NaN:

```
df_empenhos.isna().any()
```

```
Órgão                   False
Número do Processo      False
Modalidade              False
Número do Contrato       True
Número do Empenho       False
Contratado/Fornecedor   False
CNPJ/CPF                False
Valor Empenhado         False
Valor Liquidado         False
Valor Anulado           False
Valor Pago              False
Valor a Pagar           False
Fonte de Recursos       False
dtype: bool
```

```
df_dados_basicos.isna().any()
```

```
Órgão                           False
Número do Processo              False
Modalidade                      False
Objeto                          False
Valor Total/Global               True
Local de Execução/Entrega        True
Data Diário Oficial             False
Protocolo                       False
Início da vigência do contrato   True
Fim da Vigência do Contrato      True
Situação                        False
dtype: bool
```

```
df_itens_processo.isna().any()
```

```
Órgão                           False
Número do Processo              False
Modalidade                      False
Item                            False
Quantidade                      False
Unidade de Medida               False
Contratato/Fornecedor           False
CNPJ/CPF                        False
Número do Contrato               True
Inicio da Vigência do Contrato   True
Fim da Vigência do Contrato      True
Valor Unitário                  False
Valor Total/Global              False
COD                             False
dtype: bool
```

Alguns campos como Numero do Contrato, que julgo ser importante para uma licitação, se encontravam com valores nulos.

Aproximadamente 14747 registros não possuem o Número de Contrato em um total de registros neste dataset: 27610.

Portanto, opte por não aplicar a remoção destes registros pois este campo não causaria impacto direto na analise subsequente.

Remoção de registros com valores com valor inválido

Durante a limpeza e analise exploratória não foi encontrado nenhum registro que necessitaria ser removido.

Foi feito um ajuste em algumas colunas para remover caractere de cifra do Real.

```
#Converter os campos decimais
df_dados_basicos["Valor Total/Global"] = df_dados_basicos["Valor Total/Global"].str.replace('R', '').str.replace('$', '').str.rep
df_empenhos["Valor Empenhado"] = df_empenhos["Valor Empenhado"].str.replace('R', '').str.replace('$', '').str.replace('.', '').st
df_empenhos["Valor Liquidado"] = df_empenhos["Valor Liquidado"].str.replace('R', '').str.replace('$', '').str.replace('.', '').st
df_empenhos["Valor Anulado"] = df_empenhos["Valor Anulado"].str.replace('R', '').str.replace('$', '').str.replace('.', '').str.re
df_empenhos["Valor Pago"] = df_empenhos["Valor Pago"].str.replace('R', '').str.replace('$', '').str.replace('.', '').str.replace(
df_empenhos["Valor a Pagar"] = df_empenhos["Valor a Pagar"].str.replace('R', '').str.replace('$', '').str.replace('.', '').str.re
df_itens_processo["Valor Unitário"] = df_itens_processo["Valor Unitário"].str.replace('R', '').str.replace('$', '').str.replace('
df_itens_processo["Valor Total/Global"] = df_itens_processo["Valor Total/Global"].str.replace('R', '').str.replace('$', '').str.r
```

Conversão dos tipos das colunas

As colunas que foram lidas do csv original foram convertidas do tipo Object para [inteiro, decimal, datetime] utilizando os comandos asdtype ou pd.to_datetime conforme imagem abaixo:

```
#Converter os campos inteiros
df_empenhos["Número do Contrato"] = df_empenhos["Número do Contrato"].astype(pd.Int32Dtype())
df_itens_processo["Número do Contrato"] = df_itens_processo["Número do Contrato"].astype(pd.Int32Dtype())
```

```
#Converter os campos decimais
df_dados_basicos["Valor Total/Global"] = df_dados_basicos["Valor Total/Global"].str.replace('R', '').str.replace('$', '').str.rep
df_empenhos["Valor Empenhado"] = df_empenhos["Valor Empenhado"].str.replace('R', '').str.replace('$', '').str.replace('.', '').st
df_empenhos["Valor Liquidado"] = df_empenhos["Valor Liquidado"].str.replace('R', '').str.replace('$', '').str.replace('.', '').st
df_empenhos["Valor Anulado"] = df_empenhos["Valor Anulado"].str.replace('R', '').str.replace('$', '').str.replace('.', '').str.re
df_empenhos["Valor Pago"] = df_empenhos["Valor Pago"].str.replace('R', '').str.replace('$', '').str.replace('.', '').str.replace(
df_empenhos["Valor a Pagar"] = df_empenhos["Valor a Pagar"].str.replace('R', '').str.replace('$', '').str.replace('.', '').str.re
df_itens_processo["Valor Unitário"] = df_itens_processo["Valor Unitário"].str.replace('R', '').str.replace('$', '').str.replace('
df_itens_processo["Valor Total/Global"] = df_itens_processo["Valor Total/Global"].str.replace('R', '').str.replace('$', '').str.r
```

```
#Converter os campos datetime
df_dados_basicos["Data Diário Oficial"] = pd.to_datetime(df_dados_basicos["Data Diário Oficial"])
df_dados_basicos["Início da vigência do contrato"] = pd.to_datetime(df_dados_basicos["Início da vigência do contrato"])
df_dados_basicos["Fim da Vigência do Contrato"] = pd.to_datetime(df_dados_basicos["Fim da Vigência do Contrato"])
df_itens_processo["Inicio da Vigência do Contrato"] = pd.to_datetime(df_itens_processo["Inicio da Vigência do Contrato"])
df_itens_processo["Fim da Vigência do Contrato"] = pd.to_datetime(df_itens_processo["Fim da Vigência do Contrato"])
```

Junção dos datasets

Para a analise que foi feita, apenas os dois dataset foram unidos (Empenhos e Dados Basicos.

A junção foi feita usando o comando merge conforme imagem abaixo:

```
#Realizando a junção dos datasets que serão utilizados
df_juncao = df_empenhos.merge(df_dados_basicos, left_on='COD', right_on='COD', suffixes=('', '_y'), how='inner')

#este metodo permite remover as colunas repetidas após a junção
df_juncao.drop(df_juncao.filter(regex='_y$').columns.tolist(),axis=1, inplace=True)
df_juncao.info()
```

Seleção das colunas que serão utilizadas

Como os datasets possuiam algumas colunas repetidas entre si, foi necessário remover as colunas repetidas após a junção. A remoção foi feita conforme imagem abaixo:

```
#este metodo permite remover as colunas repetidas após a junção
df_juncao.drop(df_juncao.filter(regex='_y$').columns.tolist(),axis=1, inplace=True)
```

Das colunas resultantes, as que foram selecionadas são as seguintes:

```
df_juncao = df_juncao[['Órgão', 'Número do Processo', 'Modalidade', 'Contratado/Fornecedor', 'CNPJ/CPF', 'Fonte de Recursos', 'Da
df_juncao.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 28482 entries, 0 to 28481
Data columns (total 14 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Órgão                     28482 non-null  object
 1   Número do Processo        28482 non-null  object
 2   Modalidade                28482 non-null  object
 3   Contratado/Fornecedor     28482 non-null  object
 4   CNPJ/CPF                  28482 non-null  object
 5   Fonte de Recursos         28482 non-null  object
 6   Data Diário Oficial       28482 non-null  datetime64[ns]
 7   Local de Execução/Entrega 28390 non-null  object
 8   Situação                  28482 non-null  object
 9   Valor Empenhado           28482 non-null  float64
 10  Valor Liquidado           28482 non-null  float64
 11  Valor Anulado             28482 non-null  float64
 12  Valor Pago                28482 non-null  float64
 13  Valor a Pagar             28482 non-null  float64
dtypes: datetime64[ns](1), float64(5), object(8)
memory usage: 3.3+ MB
```

Conversão de valores

Algumas colunas tiveram suas informações convertidas em caixa baixa (lower case)

conforme mostra a imagem abaixo:

```
#Para evitar que um mesmo registro não entre no mesmo grupo, estarei transformando algumas colunas em Lower Case
df_juncao['Modalidade'] = df_juncao['Modalidade'].str.lower()
df_juncao['Contratado/Fornecedor'] = df_juncao['Contratado/Fornecedor'].str.lower()
df_juncao['Fonte de Recursos'] = df_juncao['Fonte de Recursos'].str.lower()
df_juncao['Local de Execução/Entrega'] = df_juncao['Local de Execução/Entrega'].str.lower()
df_juncao['Situação'] = df_juncao['Situação'].str.lower()
```

Remoção dos outliers

Durante a limpeza e analise exploratória não foi encontrado nenhum registro que necessitaria

ser removido.

Salvar arquivo limpo

O arquivo limpo foi salvo utilizando a função to_csv.

```
out_file = './data/2021-04-24_-_Base_de_Dados_Licitacoes-limpo.csv'

df_juncao.to_csv(out_file, index=False,  sep=';', encoding='UTF-8')
```

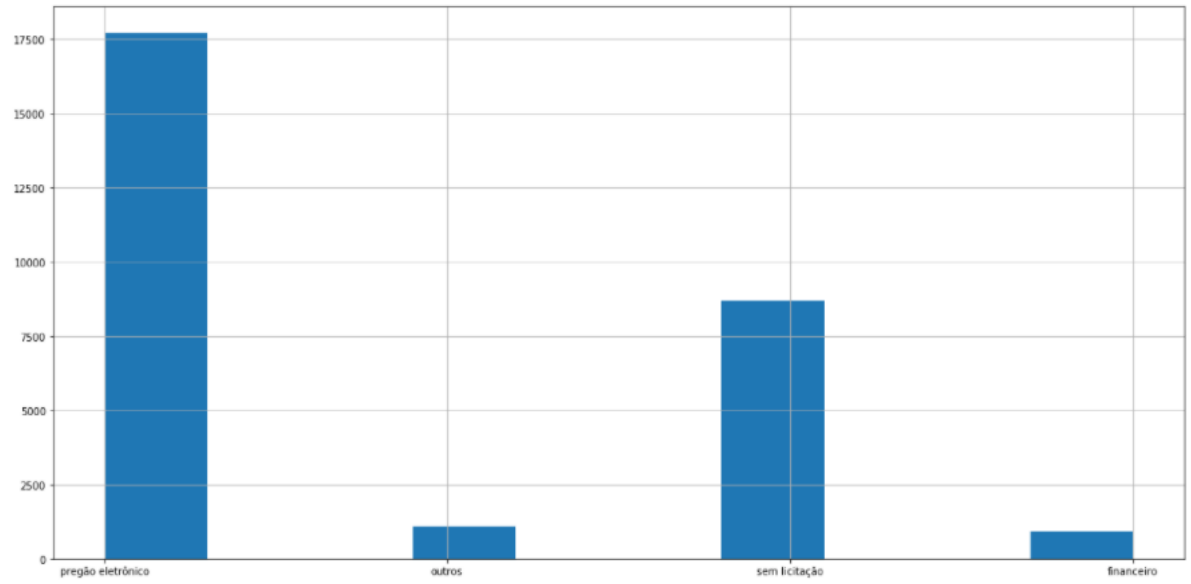
Análise Exploratória

A análise exploratória dos dados é feita da seguinte forma:

- Dimensões do dataset
- Visualização de alguns registros
- Análise da distribuição de valores
- Análise de Modalidades com base na quantidade de licitações
- Análise de Modalidades com base no valor de licitações
- Análise temporal de Modalidades com base na quantidade de licitações
- Análise temporal de Modalidades com base no valor de licitações
- Análise de outliers com base no valor de licitações
- Análise dos fornecedores com base na quantidade de licitações
- Análise dos fornecedores com base no valor de licitações
- Análise de similaridades e correlações

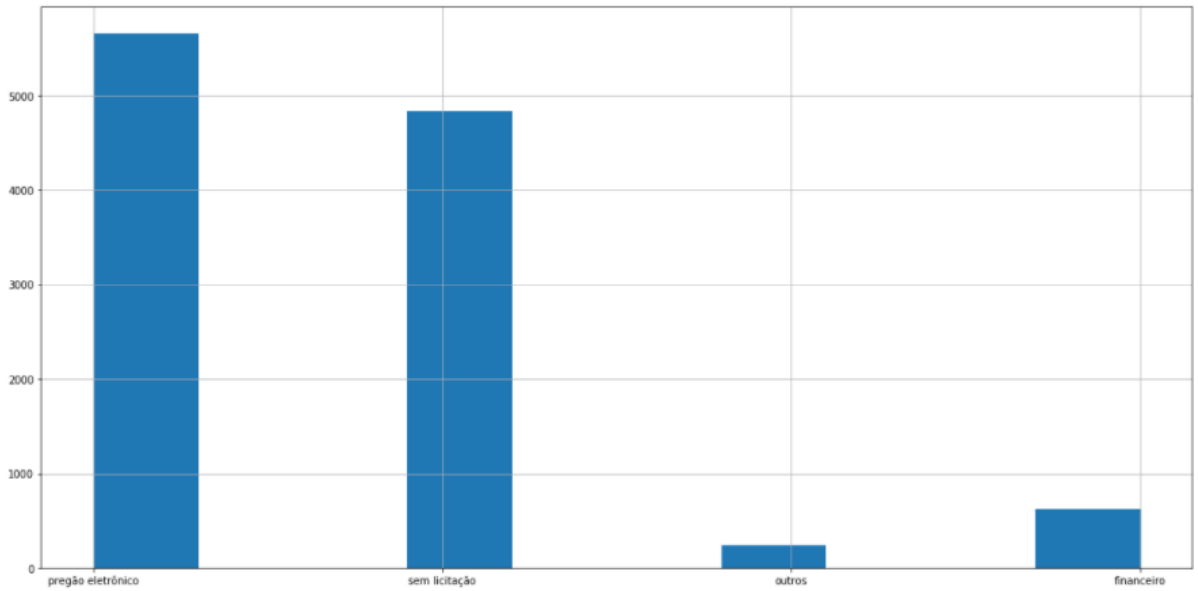
Neste relatório iremos apresenta os gráficos e análises estatísticas elaborados para cada elemento incluindo: descrição dos padrões que puderam ser observados e sugerindo algumas hipóteses para explicá-los.

Análise da distribuição de valores



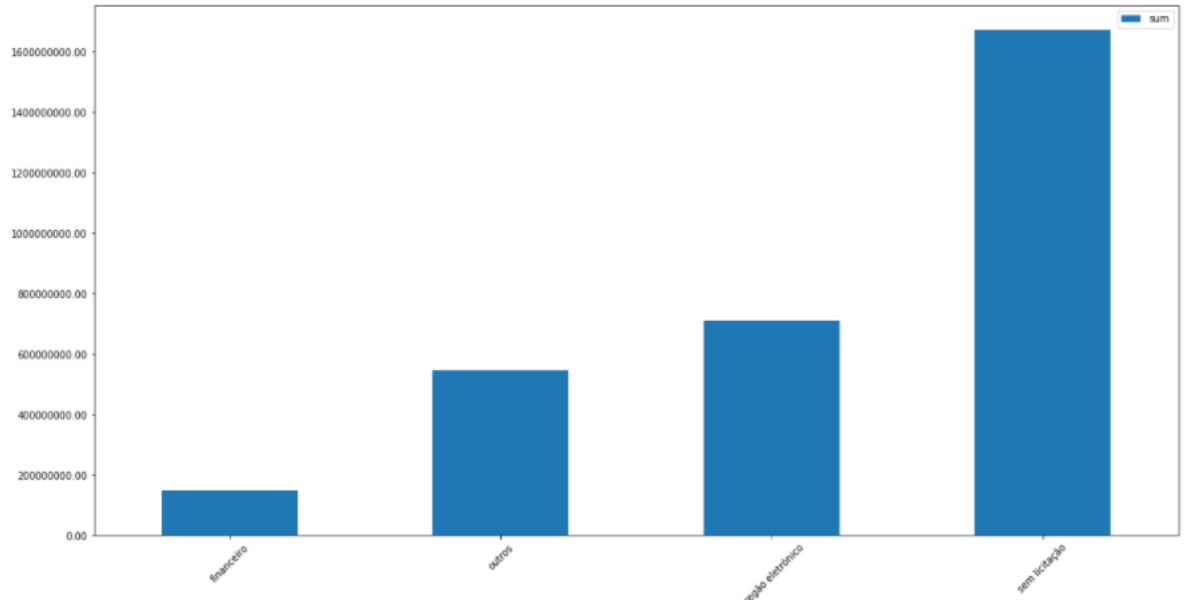
Com base no histograma do dataset, foi verificado que a maior quantidade de registros é referente à modalidade ‘Pregão Eletrônico’.

No entanto, ao comparar o histograma do periodo da pandemia covid-19, nota-se um incremento da modalidade “Sem Licitação”.



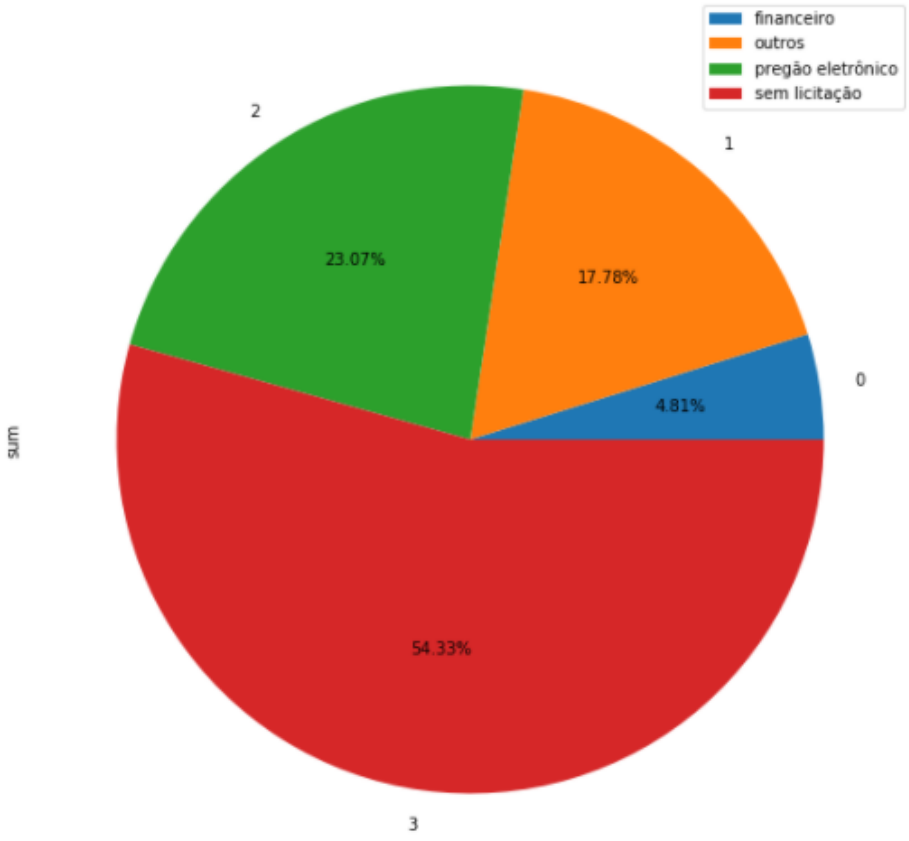
A hipotese principal é que a pandemia possibilitou que a prefeitura obtivesse serviços e produtos sem licitação devido as caracteristicas emergenciais que este periodo nos proporcionou.

Análise de Modalidades com base no valor de licitações

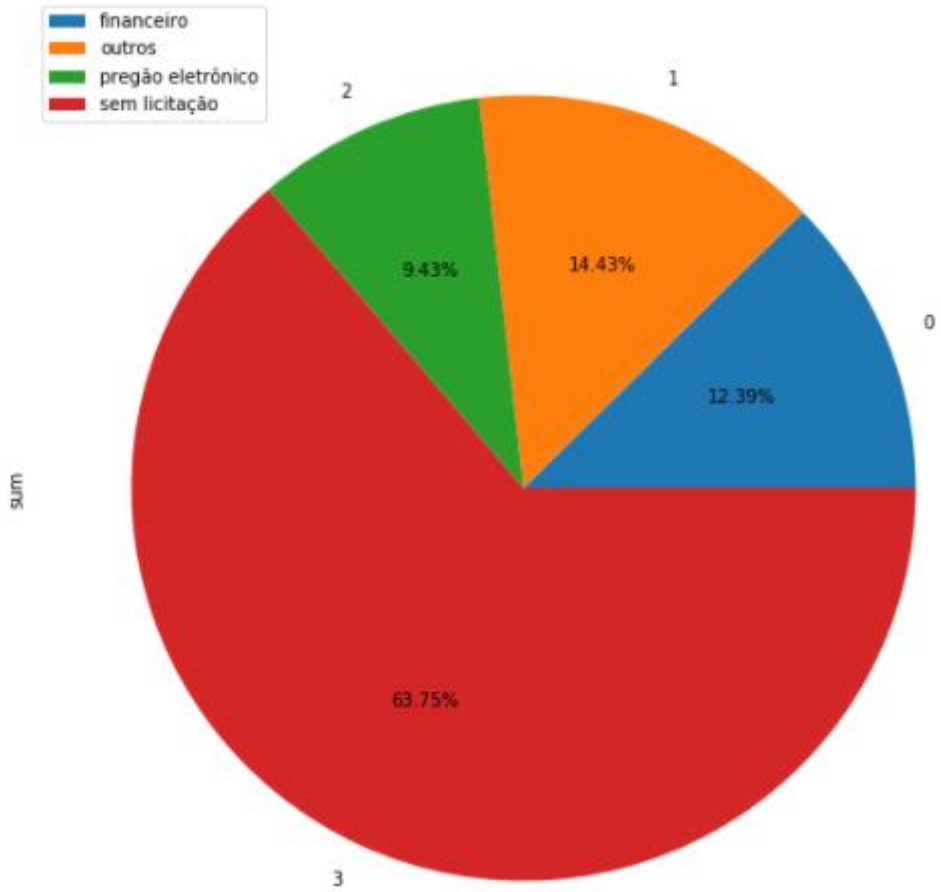


Com base neste grafico, é possivel notar que a soma mensal dos valores de licitações é maior para a modalidade “Sem Licitação”.

Sendo este, cerca de 54% do montante total conforme mostra o pie chart abaixo:

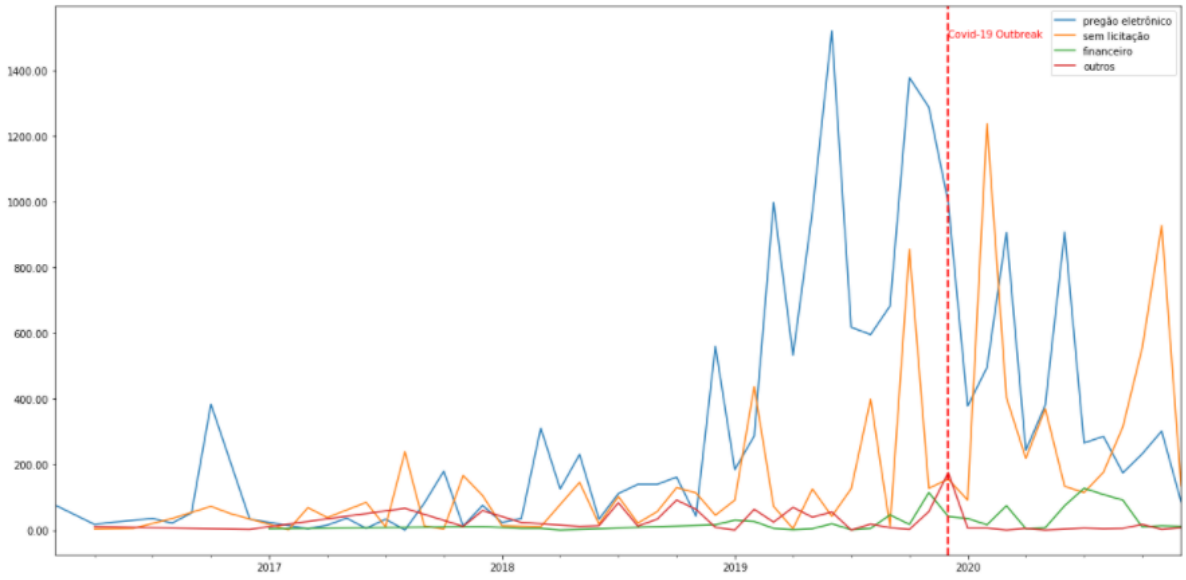


Comparando com o período da pandemia, observa-se um aumento de licitações sem licitação.



Conforme visualizado anteriormente, este incremento de cerca de 10% pode ser explicado pela quantidade de modalidade “Sem licitação” que também aumentou.

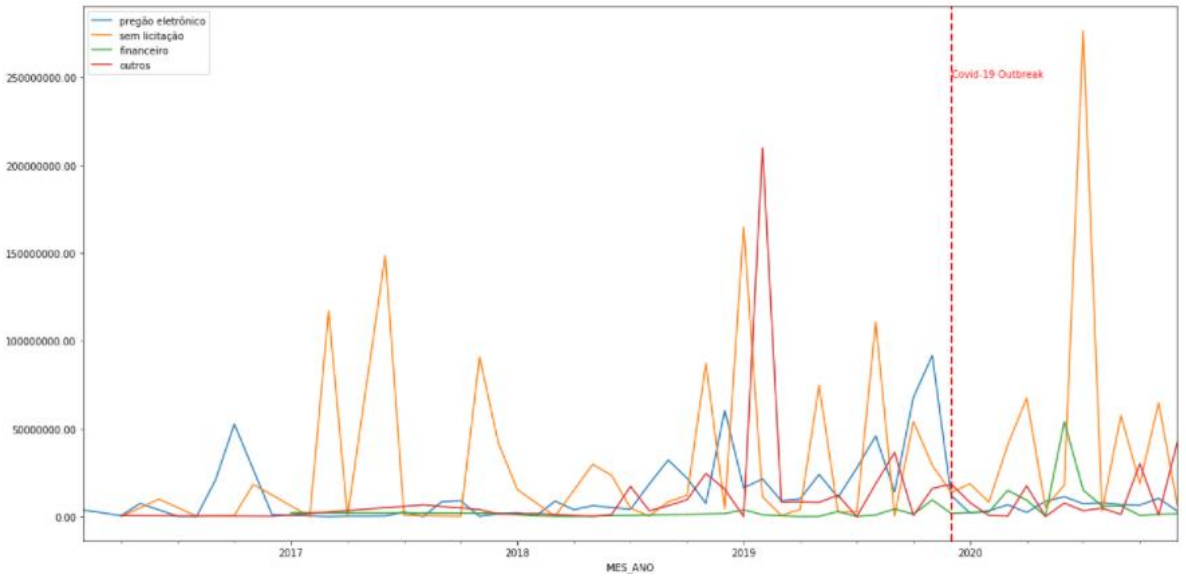
Análise temporal de Modalidades com base no quantidade de licitações



A quantirade de licitações da modalidade “Sem licitação” parece ter aumentado significativamente um pouco antes do periodo da pandemia e continuou alto até o fim dos registros deste dataset.

Observa-se também que a quantidade de licitações da modalidade “Pregão eletrônico” caiu proporcionalmente.

Análise temporal de Modalidades com base no valor de licitações

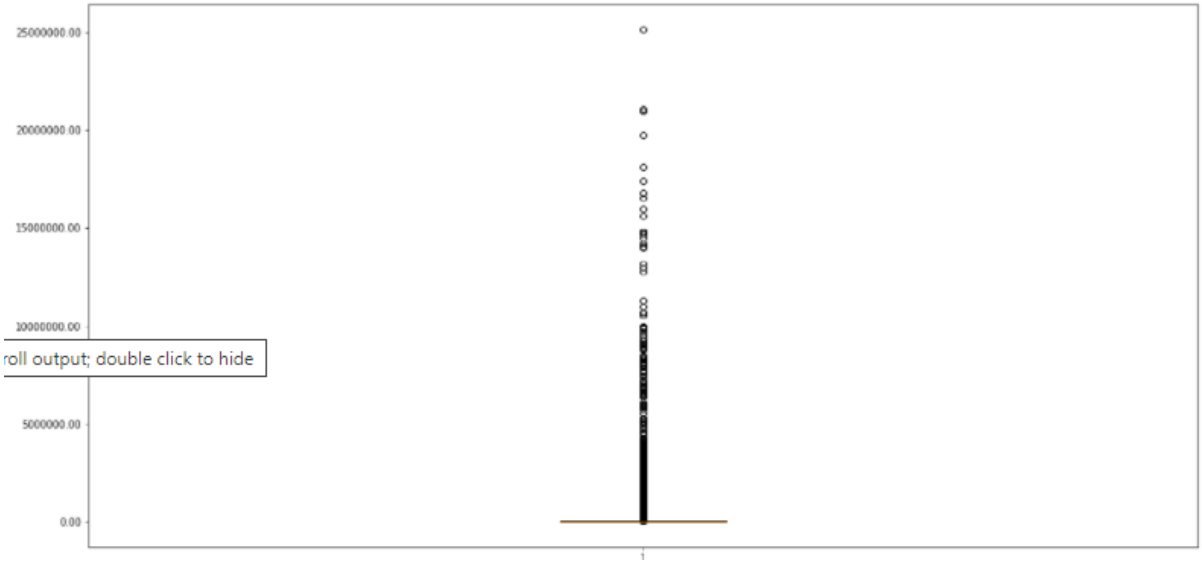


Neste grafico, nota-se que o valor total mensal sem licitação teve um aumento momentaneo após a pandemia e depois se manteve praticamente constante.

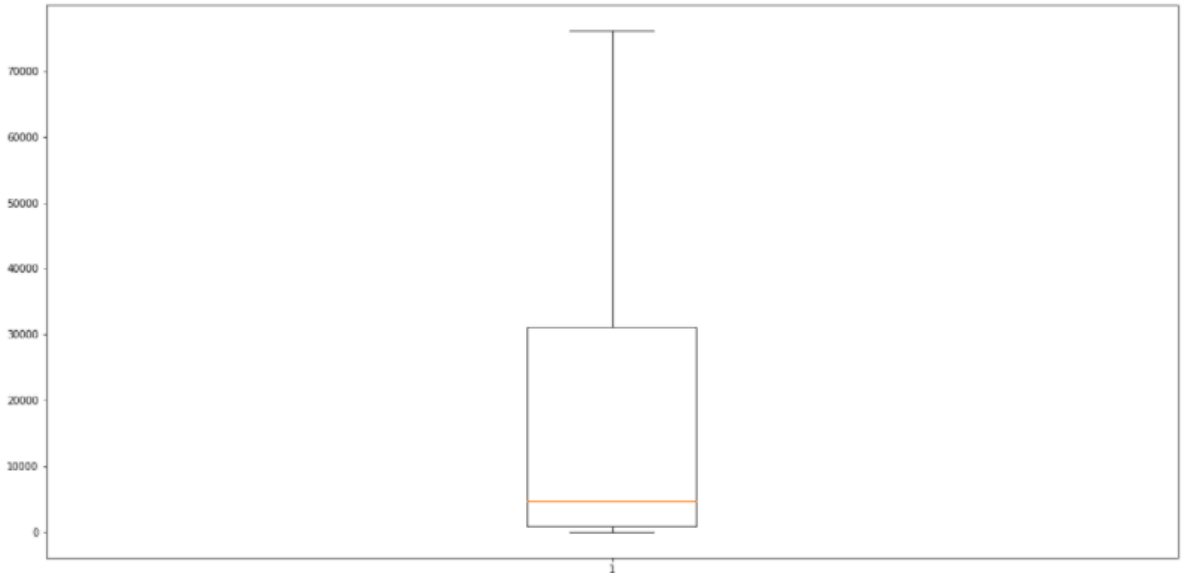
A hipotese é que este pico na metade do ano de 2020 pode ser explicado por um repasse de verbas do governo federal nos estados e cidades para combate ao Covid-19.

Análise de outliers com base no valor de licitações

Ao explorar o boxplot dos valores de licitações, foi possivel contatar que muitos valores são considerados outliers.



Visto que boa parte dos registros estão localizados nesta faixa de valores:

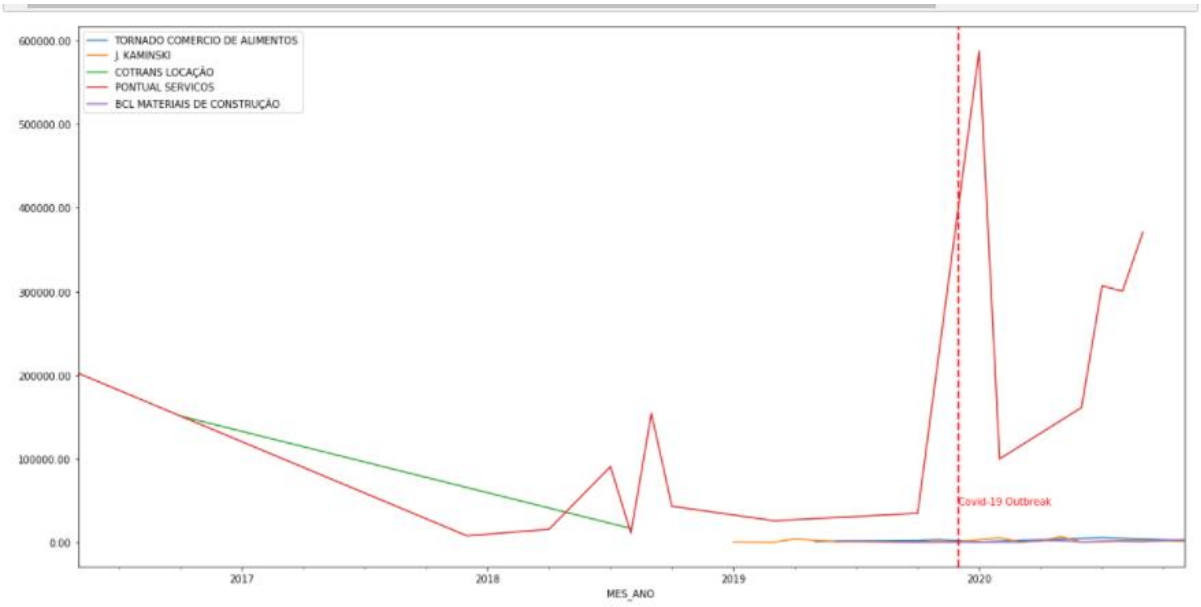


A hipotese disto acontecer é que a prefeitura costuma fazer contratos grandes e de longe duração, gerando alguns periodos com pouca movimentação de licitações.

Análise dos fornecedores com base na quantidade e valor de licitações

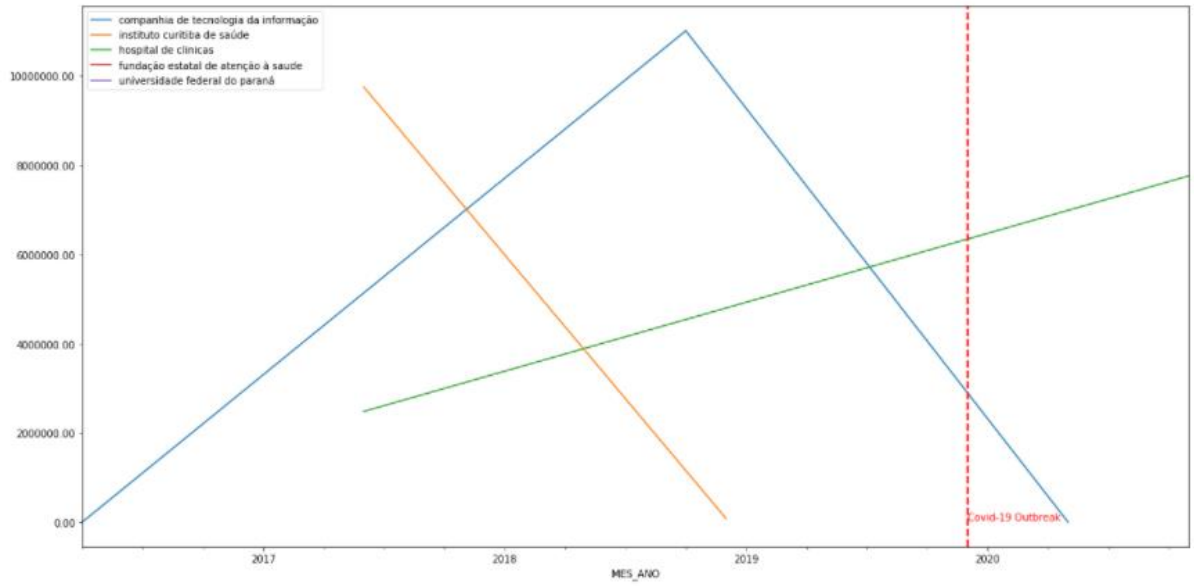
Feito feito uma analise dos fornecedores com base na quantidade e valor total das licitações.

Dentre todos os fornecedores, os 5 maiores foram analisados nos graficos abaixo:



Observa-se que o fornecedor de alimentos “Tornado Comercio de Alimentos” obteve um aumento médio de valores em licitações.

No entanto, no proximo grafico, quando comparamos os gastos médios sem licitação, verificamos que boa parte dos fornecedores são categorizados como serviços de saúde.



Análise de similaridades e correlações

Algumas analises foram realizadas para verificar similaridades entre os campos do dataset.

```
df_similaridades = pd.crosstab(df_juncao['Modalidade'], df_juncao['Situação'])
df_similaridades
```

Modalidade \ Situação	confirmado vencedor	cotação informada	empenhado	parcialmente empenhado	pedido ou empenho anulado	processo concluído
financeiro	948	0	0	0	0	0
outros	1017	0	7	17	0	51
pregão eletrônico	4035	0	3067	10634	0	0
sem licitação	8032	1	475	197	1	0

Situação	confirmado vencedor	cotação informada	empenhado	parcialmente empenhado	pedido ou empenho anulado	processo concluído
confirmado vencedor	1.000000	0.902506	0.254292	0.121858	0.902506	-0.496937
cotação informada	0.902506	1.000000	-0.186953	-0.317425	1.000000	-0.333333
empenhado	0.254292	-0.186953	1.000000	0.990907	-0.186953	-0.399187
parcialmente empenhado	0.121858	-0.317425	0.990907	1.000000	-0.317425	-0.340143
pedido ou empenho anulado	0.902506	1.000000	-0.186953	-0.317425	1.000000	-0.333333
processo concluído	-0.496937	-0.333333	-0.399187	-0.340143	-0.333333	1.000000

Perguntas iniciais

Nesta seção iremos listar algumas perguntas as quais gostariamos de ter as respostas à partir das analises obtidas neste trabalho.

- Houveram alterações significa nas licitações comparando o antes e o depois da pandemia do Covid-19?
- Houve aumento na quantidade de contratações sem licitação durante a pandemia?
- Houve aumento nos gastos sem licitação durante a pandemia?

Referências

Prefeitura de Curitiba. *Dados Abertos*. Disponível em:

<https://www.curitiba.pr.gov.br/DADOSABERTOS/>. Acesso em: 25 abril. 2021.

Ministérios da Saúde. *Dados Corona Vírus*. Disponível em:

<https://coronavirus.saude.gov.br/linha-do-tempo/>. Acesso em: 25 abril. 2021.