Calidad e Integración de Datos

Diego Velasco
Federico Carneiro

Introducción

El trabajo tiene como finalidad el análisis e implementación del modelo de Calidad aplicando todas la definiciones trabajadas en el curso. (ver letra_proyecto_parte1 y letra_proyecto_parte_2)
En el documento se pretendió detallar todas las determinaciones que fuimos tomando de acuerdo a los conceptos vistos en clase junto con las definiciones de cada métrica y su justificación. También se incluyó la instancia de cada una de ellas a los datos en cuestión según su granularidad.

En las siguientes diapositivas resumimos el trabajo realizado en la elaboración del modelo de calidad para los archivos presentados para el trabajo final del curso.

Un análisis de los datos propuestos, Profiling de los mismos y cuál fue la herramienta que utilizamos y algunos mecanismos, resumen de las métricas y sus instancias, tanto para los archivos estructurados como para el que no lo es, y finalizando con el modelo propuesto para almacenar las mediciones y la incorporación del contexto en el Modelo de Calidad Inicial.

Datos

emisivos.csv

	A	B	C	D	E	F	G
1	IdLugarSalid	Lugar Salida	IdTranspSali	Transporte Ir	FechaSalida	IdFecSalida	FechaEr
2	4	Chuy	2	Terrestre Au	29/12/2016	13512	10/1/
3	18		2	Terrestre Au	29/12/2016	13512	9/1/
4	3	Aeropuerto (	1	Aereo	3/1/2017	13517	27/1/
5	3	Aeropuerto (	1	Aereo	6/1/2017	13520	21/1/
6	4	Chuy	2	Terrestre Au	2/1/2017	13516	25/1/
7	18	Río Branco	2	Terrestre Au	30/12/2016	13513	9/1/
8	7	Fray Bentos	3	Terrestre Bu	30/12/2016	13513	3/1/
9	16	Bella Unión	2	Terrestre Au	5/1/2017	13519	11/1/
10	4	Chuy	2	Terrestre Au	30/12/2016	13513	9/1/
11	4	Chuy	3	Terrestre Bu	18/12/2016	13501	7/1/
12	18	Río Branco	2	Terrestre Au	10/1/2017	13524	16/1/
13	7	Fray Bentos	3	Terrestre Bu	19/1/2017	13533	25/1/
14	1	Colonia	5	Maritimo - Fl	6/1/2017	13520	9/1/
15	4	Chuy	3	Terrestre Bu	27/12/2016	13510	7/1/
16	1	Colonia	5	Maritimo - Fl	10/1/2017	13524	16/1/
17	10	Salto	2	Terrestre Au	01/27/2017	13541	30/1/
18	8	Paysandú	3	Terrestre Bu	01/17/2017	13531	22/1/
19	8		2	Terrestre Au	13/1/2017	13527	16/1/

operadores.csv

TipoOperado	Operador	Departamen	Localidad	Direccion	Telefono
Inmobiliaria	DESTINO PUI	MALDONADO	MALDONADO	SIMON BOLI	950023
Inmobiliaria	S/D	ROCHA	CHUY	1º DE AGOST	985746
Inmobiliaria	MONICA PUC	MALDONADO	PUNTA BALL	RUTA INTERE	944332
Inmobiliaria	SPOT REAL E	MONTEVIDE	MONTEVIDE	POTOSI 1657	260571
Inmobiliaria	SANITAGO P	MALDONADO	PUNTA DEL E	AV FRANCIS	424850
Inmobiliaria	BON PORTO	ROCHA	LA PALOMA	MARACOPA	S/D
Alojamiento	HOTEL SORO	MONTEVIDE	MONTEVIDE	FEDERICO N.	271080
Alojamiento	GRAN HOTEL	SALTO	SALTO	URUGUAY 74	996159
Inmobiliaria	INMOBILIAR	COLONIA	COLONIA DE	18 DE JULIO 2	452297
Turismo ave	SENDERO CU	RIVERA	RIVERA	RUTA 30 KM	990762
SALAS DE CO	HOTEL PUNT	MONTEVIDE	MONTEVIDE	S/D	271209
SALAS DE CO	XENIA	ARTIGAS	BELLA UNION	AV. ARTIGAS	47794848
Inmobiliaria	DIONI STUAF	MALDONADO	MALDONADO	BATLLE Y ORI	422454
Inmobiliaria	GONZALEZ B	CANELONES	CIUDAD DE L	FILADELFIA M	973495
Agencia de t	LAS VEGAS	ARTIGAS	BELLA UNION	ING. ALFRED	477952
SALAS DE CO	HOTEL COST	MONTEVIDE	MONTEVIDE	PLAZA INDEF	270686
SALAS DE CO	DUTY FREE A	RIVERA	RIVERA	21 DE SETIEM	241223
SALAS DE CO	PROVIMAR L	PAYSANDU	PAYSANDU	AVDA. GRAL.	260538

receptivos.json


```
[
  {
    "IdIngresos": 3,
    "Lugar Ingreso": "Aeropuerto de Carras
    "IdTranspIngreso": 1,
    "Transporte Internacional de Ingreso":
    "FechaIngreso": "2017-02-22",
    "IdFecIng": 13567,
    "FechaEgreso": "2017-03-03",
    "IdFecEgr": 13576,
    "IdNacionalidad": 33,
    "Pais": "Ecuador",
    "IdResidencia": 50,
    "Residencia": "Otras ciudades Sudameri
    "IdMotivo": 99,
    "Motivo": "Otros",
    "IdOcupacion": 20,
    "Ocupacion": "Deportista, Entrenador,
    "IsEstudio": 4,
    "Estudio": "Secundaria completa",
```

Datos

emisivos.csv y operadores.csv

Los archivos de Turismo emisivo y Operadores turísticos se encuentran almacenados en archivos csv los cuales se encuentran fuertemente estructurados. Esto permite inferir algunas estructuras y tipos de datos a partir sus datos.

receptivos.json

El archivo de Turismo receptivo se encuentra almacenado en formato json. Es un formato desestructurado, donde su contenido puede llegar a ser muy variado en cuanto a su estructura y sus anidaciones.

emisivos.csv	
Cantidad de Registros (filas)	20602
Cantidad de Columnas	43
Cantidad de Columnas Enteras	14
Cantidad de Columnas Strings	14
Cantidad de Columnas Decimales	15

operadores.csv	
Cantidad de Registros (filas)	3288
Cantidad de Columnas	10
Cantidad de Columnas Enteras	0
Cantidad de Columnas Strings	8
Cantidad de Columnas Decimales	0
Cantidad de Columnas Coordenadas Geográficas	2

receptivos.json	
Cantidad de Registros (filas)	47785
Cantidad de Columnas	48
Cantidad de Columnas Enteras	19
Cantidad de Columnas Strings	17
Cantidad de Columnas Decimales	10
Cantidad de Columnas Date	2

Data Profiling

emisivos.csv

Alerts

`Lugar Salida` has 1993 (9.7%) missing values	Missing
`IdDeptoResidencia` has 2014 (9.8%) missing values	Missing
`Departamento` has 2014 (9.8%) missing values	Missing
`IdMotivo` has 610 (3.0%) missing values	Missing
`Motivo` has 610 (3.0%) missing values	Missing
`GastoAlimentacion` is highly skewed (γ1 = 26.52016254)	Skewed
`IdTranspLocal` has 5174 (25.1%) zeros	Zeros
`GastoTotal` has 221 (1.1%) zeros	Zeros
	Zeros
	Zeros
	Zeros
	Zeros
	Zeros
	Zeros
	Zeros
	Zeros

Overview | Alerts 16 | Reproduction

Dataset statistics

Number of variables	43
Number of observations	20602
Missing cells	7241
Missing cells (%)	0.8%
Total size in memory	6.8 MiB
Average record size in memory	344.0 B

Variable types

Numeric	29
Text	14

Lugar Salida

Text

`MISSING`

Distinct	15
Distinct (%)	0.1%
Missing	1993
Missing (%)	9.7%
Memory size	161.1 KiB

Data Profiling

operadores.csv

Dataset statistics

Number of variables	10
Number of observations	3288
Missing cells	1
Missing cells (%)	< 0.1%
Total size in memory	257.0 KiB
Average record size in memory	80.0 B

Variable types

Text	10

TipoOperador
Text

Distinct	12
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	25.8 KiB

```
+------------------+--------+------------+-------------+----------------+---------+---+------------------+--------+-------+
|      TipoOperador|Operador|Departamento|    Localidad|       Direccion| Telefono|Web|             EMail|Longitud|Latitud|
+------------------+--------+------------+-------------+----------------+---------+---+------------------+--------+-------+
|Rentadora de autos|EUROPCAR|   CANELONES|PASO CARRASCO|26 DE AGOSTO 610|094404570|S/D|silbus@vera.com.uy|  -10000| -10000|
|Rentadora de autos|EUROPCAR|   CANELONES|PASO CARRASCO|26 DE AGOSTO 610|094404570|S/D|silbus@vera.com.uy|  -10000| -10000|
|Rentadora de autos|EUROPCAR|   CANELONES|PASO CARRASCO|26 DE AGOSTO 610|094404570|S/D|silbus@vera.com.uy|  -10000| -10000|
|Rentadora de autos|EUROPCAR|   CANELONES|PASO CARRASCO|26 DE AGOSTO 610|094404570|S/D|silbus@vera.com.uy|  -10000| -10000|
+------------------+--------+------------+-------------+----------------+---------+---+------------------+--------+-------+
```

Data Profiling

receptivos.json

Overview | Alerts 30 | Reproduction

Dataset statistics

Number of variables	48
Number of observations	47785
Missing cells	0
Missing cells (%)	0.0%
Total size in memory	17.5 MiB
Average record size in memory	384.0 B

Variable types

Unsupported	30
Text	18

Lugar Ingreso
Text

Distinct	16
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
	373.4 KiB

```
+--------------+-------------+-----------------------+--------------+-------------------------+-----------------+
|IsEstudio_nulos|Estudio_nulos|IdDestinoLocalidad_nulos|Localidad_nulos|IdDepartamentoDestino_nulos|Departamento_nulos|
+--------------+-------------+-----------------------+--------------+-------------------------+-----------------+
|         18393|        18393|                  15490|         15490|                    15490|            15490|
+--------------+-------------+-----------------------+--------------+-------------------------+-----------------+
```

```
+--------------+-----------------+-----------------+-----------------+-----------------------------------+----------------+--------------+----------------+
|IdIngresos_nulos|Lugar_Ingreso_nulos|IdTranspIngreso_nulos|Transporte_Internacional_Ingreso_nulos|FechaIngreso_nulos|IdFecIng_nulos|FechaEgreso_nulos|
+--------------+-----------------+-----------------+-----------------+-----------------------------------+----------------+--------------+----------------+
|         45286|            45286|            45286|                                45286|           45286|           45286|           45286|
+--------------+-----------------+-----------------+-----------------+-----------------------------------+----------------+--------------+----------------+
```

```
+--------------+---------------------+----------+------------------+---------------+--------------+-----------+----------------+-------------+
|IdFecEgr_nulos|IdNacionalidad_nulos|Pais_nulos|IdResidencia_nulos|Residencia_nulos|IdMotivo_nulos|Motivo_nulos|IdOcupacion_nulos|Ocupacion_nulos|
+--------------+---------------------+----------+------------------+---------------+--------------+-----------+----------------+-------------+
|         45286|                45286|     45286|             45286|          45286|         45286|      45286|           45286|        45286|
+--------------+---------------------+----------+------------------+---------------+--------------+-----------+----------------+-------------+
```

Métricas de Calidad

Modelo Relacional

Exactitud

Exact_sintactica_bool	
Descripción	Evalúa la correctitud sintáctica de una celda
Unidades	{0,1}
Granularidad	Celda

Exact_sintactica_dist	
Descripción	Evalúa la correctitud sintáctica de una celda
Unidades	[0..1]
Granularidad	Celda

Exact_semantica_bool	
Descripción	Evalúa la correctitud semántica de una celda
Unidades	{0,1}
Granularidad	Celda

Exact_precision	
Descripción	Evalúa la precisión de una celda
Unidades	{0, 0.33, 0.66, 1}
Granularidad	Celda

Completitud

Completitud_cobertura	
Descripción	Mide si una tabla cubre todos los registros que debería cubrir.
Unidades	[0..1]
Granularidad	Tabla

Completitud_densidad	
Descripción	Calcula la ratio de valores faltantes en una columna con respecto a la cantidad total de valores
Unidades	[0..1]
Granularidad	Columna

Frescura

ratio_actualidad	
Descripción	Evalúa la actualidad de la tabla
Unidades	[0..1]
Granularidad	Tabla

Consistencia

Consistencia_Intra_Relacion	
Descripción	Evalúa la consistencia entre dos datos distintos
Unidades	{0,1}
Granularidad	Celda

Consistencia_Intra_Relacion_columna	
Descripción	Evalúa la consistencia entre dos columnas distintas
Unidades	{0,1}
Granularidad	Columna

Consistencia_Intra_Relacion_conj_celdas	
Descripción	Evalúa la consistencia entre un conjunto de celdas
Unidades	{0,1}
Granularidad	Conjunto de celdas

Consistencia_Int_Dominio	
Descripción	Evalúa la integridad de dominio de un dato
Unidades	{0,1}
Granularidad	Celda

Unicidad

Unicidad_no_duplicacion_tupla	
Descripción	Evalúa si un registro es un duplicado exacto o no
Unidades	{0,1}
Granularidad	Tupla

Unicidad_no_contradiccion_tupla	
Descripción	Evalúa si un registro es un duplicado contradictorio o no
Unidades	{0,1}
Granularidad	Tupla

Métricas Instanciadas

Modelo Relacional

Exactitud

Exact_sintactica_bool_FechaSalida	
Métrica	Exact_sintactica_bool
Datos	emisivos.'FechaSalida'
Método Medición	If check_date_format(FechaSalida) == True exactitud = 1 else: exactitud = 0

Exact_sintactica_dist_direccion	
Métrica	Exact_sintactica_dist
Datos	operadores.Direccion
Método Medición	exactitud = 1 - min_levenshtein_distance(Direccion, referencial)

Exact_semántica_bool_`Lugar Salida`	
Métrica	Exact_semantica_bool
Datos	emisivos.`Lugar Salida`
Método Medición	if `Lugar Salida` not in list_lugares_salida exactitud = 0 else: exactitud = 1

Exact_precision_Destino	
Métrica	Exact_precision
Datos	emisivos.Destino

Método Medición	if Destino is pais:
	exactitud = 1.00
	elif Destino is region:
	exactitud = 0.66
	elif Destino is continente:
	exactitud = 0.33
	else:
	exactitud = 0

Completitud

Completitud_cobertura_emisivos	
Métrica	Completitud_densidad
Datos	emisivos
Método Medición	cobertura = len(tabla) / L_ref_estim

Completitud_densidad_Lugar Salida	
Métrica	Completitud_densidad
Datos	emisivos.Lugar Salida
Método Medición	densidad = suma_nulos(df['Lugar Salida'])

Frescura

ratio_actualidad_emisivos	
Métrica	ratio_actualidad
Datos	emisivos
Método Medición	frescura = 1 - (t2-t0) / Δt0

Consistencia

Consistencia_Intra_Relacion_Lugar.Salida	
Métrica	Consistencia_Intra_Relacion
Datos	emisivos.'Lugar Salida'

Método Medición	if 'Lugar Salida'_!= dict_lugares_salida[IdLugarSalida]:
	consistencia = 0
	else:
	consistencia = 1

Consistencia-Intra-Relacion-Lugar.Salida	
Métrica	Consistencia_Intra_Relacion_columna
Datos	emisivos.'Lugar Salida'
Método Medición	if set(df['Lugar Salida'].unique()) != set(df['Lugar Ingreso'].unique()) consistencia = 0 else: consistencia = 1

Consistencia_Intra_Relacion_Fechas_Estadia	
Métrica	Consistencia_Intra_Relacion_conj_celdas
Datos	emisivos.FechaEntrada, emisivos.FechaSalida, emisivos.Estadia
Método Medición	if FechaEntrada - FechaSalida != Estadia consistencia = 0 else: consistencia = 1

Consistencia_Int_Dominio_Latitud	
Métrica	Consistencia_Int_Dominio
Datos	emisivos.latitud
Método Medición	consistencia = dist_lat_norm (latitud, departamento)

Unicidad

Unicidad_no_duplicacion_tupla_emisivos	
Métrica	Unicidad_emisivos_no_duplicacion
Datos	emisivos.IdLugarSalida, emisivos.IdTranspSalidad, emisivos.Transporte Internacional de Salida, emisivos.IdFecSalida, emisivos.FechaEntrada, emisivos.IdFecEntrada, emisivos.IdNacionalidad emisivos.Pais, emisivos.IdDeptoResidencia, emisivos.Departamento, emisivos.IdMotivo emisivos.Motivo, emisivos.IdOcupacion, emisivos.Ocupacion emisivos.IdNivelEstudio, emisivos.Estudio, emisivos.IdDestino, emisivos.Destino, emisivos.IdAlojamiento, emisivos.Alojamiento emisivos.IdLugarIngreso, emisivos.Lugar Ingreso emisivos.IdTranspIngreso, emisivos.Transporte Internacional de Ingreso, emisivos.IdTranspLocal, emisivos.Trasporte Local emisivos.Estadia, emisivos.Gente, emisivos.GastoTotal, emisivos.GastoAlojamiento, emisivos.GastoAlimentacion emisivos.GastoTransporteInternac, emisivos.GatoTransporteLocal, emisivos.GastoCultural, emisivos.GastoTours, emisivos.GastoCompras, emisivos.GastoResto
Método Medición	unicidad = duplicados_row(indice)['no_duplicacion']

Unicidad_no_duplicacion_tupla_emisivos	
Métrica	Unicidad_emisivos_no_contradiccion
Datos	emisivos.IdLugarSalida, emisivos.IdTranspSalidad, emisivos.Transporte Internacional de Salida, emisivos.IdFecSalida, emisivos.FechaEntrada, emisivos.IdFecEntrada, emisivos.IdNacionalidad emisivos.Pais, emisivos.IdDeptoResidencia, emisivos.Departamento, emisivos.IdMotivo emisivos.Motivo, emisivos.IdOcupacion, emisivos.Ocupacion emisivos.IdNivelEstudio, emisivos.Estudio, emisivos.IdDestino, emisivos.Destino, emisivos.IdAlojamiento, emisivos.Alojamiento emisivos.IdLugarIngreso, emisivos.Lugar Ingreso emisivos.IdTranspIngreso, emisivos.Transporte Internacional de Ingreso, emisivos.IdTranspLocal, emisivos.Trasporte Local emisivos.Estadia, emisivos.Gente, emisivos.GastoTotal, emisivos.GastoAlojamiento, emisivos.GastoAlimentacion emisivos.GastoTransporteInternac, emisivos.GatoTransporteLocal, emisivos.GastoCultural, emisivos.GastoTours, emisivos.GastoCompras, emisivos.GastoResto
Método Medición	unicidad = duplicados_row(indice)['no_contradiccion']

Métricas Instanciadas

Modelo Relacional - Métodos de medición

Exactitud

```python
def check_date_format(string):
    pattern = r"\d{4}-\d{2}-\d{2}"
    match = re.fullmatch(pattern, string)
    if match:
        return True
    else:
        return False
```

```python
import Levenshtein

def min_levenshtein_distance(string, referencial):
    # Se sacan los números de la columna de Dirección
    modified_string = ''.join(char for char in string if not char.isdigit())

    # Se calcula la mínima distancia de Levenshtein a algún valor del referencial
    # y se guarda el valor
    min_distance = float('inf')
    min_value = None

    for value in referencial:
        distance = Levenshtein.distance(modified_string, str(value))
        if distance < min_distance:
            min_distance = distance
            min_value = value

    return round(min_distance/(len(modified_string) + len(min_value)),4)
```

Completitud

```python
def suma_nulos(column):
    null_count = column.isnull().sum()
    sin_datos_count = (column == "Sin Datos").sum()
    desc_sin_datos_count = (column == "Desconocido / Sin Datos").sum()
    s_d_count = (column == "S/D").sum()

    cant_nulos = null_count + sin_datos_count + desc_sin_datos_count + s_d_count
    return round(cant_nulos/len(column),4)
```

Dic 2016	47.624	No hay datos, se estima como dic-21
2017	1.788.792	
2018	1.788.792	No hay datos, se estima como 2017
2019	2.199.152	
2020	599.512	1er trimestre (hasta comienzo de pandemia)
2021	92.517	Noviembre y diciembre ("fin de pandemia")
2022	2.383.901	
2023	1.141.301	1er trimestre
Tot	**10.041.591**	**Valor asignado para L_ref_estim**

Consistencia

```python
def dist_lat_norm (latitud, departamento):
    max_lat = dict_coord[departamento]['max_lat']
    min_lat = dict_coord[departamento]['min_lat']
    if latitud > max_lat:
        distancia_norm = abs(latitud - max_lat) / abs(max_lat - (90))
    elif latitud < min_lat:
        distancia_norm = abs(latitud - min_lat) / abs(min_lat - (-90))
    else:
        distancia_norm = 0

    return round(distancia_norm,4)
```

Métricas Instanciadas

Modelo Relacional - Métodos de medición

Unicidad

```python
def duplicados_row (df,row):
    subset_columns = list(df.columns)
    cols_not_to_consider = ['Coef', 'CoefTot', 'latitud', 'longitud', 'FechaSalida', 'Lugar Salida']

    # Genero listado de columnas a partir de las cuales se buscan duplicados
    for col in cols_not_to_consider:
        subset_columns.pop(subset_columns.index(col))

    # Genero dataframe con duplicados
    df_duplicates = df[df.duplicated(subset=subset_columns,keep=False)].sort_values(by = subset_columns)

    df_dup_row = df_duplicates[(df_duplicates[subset_columns] == df.loc[row][subset_columns]).all(axis=1)] # Dataframe con duplicados para cada registro duplicado

    if len(df_dup_row) == 0:
        no_duplicacion = 1
        no_contradiccion = 1
        return {'no_duplicacion': no_duplicacion, 'no_contradiccion': no_contradiccion}

    # Genero listas con los atributos que definiran si el duplicado es contradictorio o no
    for col in cols_not_to_consider:
        coef_list = [coef for coef in df_dup_row['Coef']]
        coef_tot_list = [coef_tot for coef_tot in df_dup_row['CoefTot']]
        lat_list = [lat for lat in df_dup_row['latitud']]
        lon_list = [lon for lon in df_dup_row['longitud']]
    check_list = [coef_list, coef_tot_list, lat_list, lon_list]

    to_remove = [None, 'Sin Datos', 'Desconocido / Sin Datos'] # valores a remover para chequear si los restantes son contradictorios

    # Chequeo si los duplicados son contradictorios
    for lista in check_list:
        contradiccion = []
        for value in to_remove:
            lista.remove(value) if value in lista else lista
        if len(lista) == len(set(lista)):
            contradiccion.append(True) # Existen valores distintos luego de sacar los nulos o faltantes
        else:
            contradiccion.append(False) # No existen valores distintos luego de sacar los nulos o faltantes

    if True in contradiccion:
        no_duplicacion = 1
        no_contradiccion = 0
    else:
        no_duplicacion = 0
        no_contradiccion = 1

    return {'no_duplicacion': no_duplicacion, 'no_contradiccion': no_contradiccion}
```

Métricas de Calidad

Modelo No Estructurado - Convención

Granularidades

- **Archivo**: Una métrica con esta granularidad toma un único valor para todo el archivo JSON.
- **Documento**: Una métrica con esta granularidad toma un único valor para cada documento.
- **Atributo**: Una métrica con esta granularidad toma un único valor para un atributo. Dicho valor comprende a todos los documentos para los cuales esté definido, o no, dicho atributo. Para instanciar una métrica con esta granularidad, se debe indicar el atributo para el cual se instancia.
- **Valor**: Una métrica con esta granularidad toma un único valor para un valor asociado a algún atributo dentro de un documento. Para instanciar una métrica con esta granularidad, se debe indicar el atributo cuyo valor será utilizado para realizar la medición.
- **Conjunto de valores**: Una métrica con esta granularidad toma un único valor para un conjunto de valores dentro de un documento. Para instanciar una métrica con esta granularidad, se debe indicar los atributos cuyos valores serán utilizados para realizar la medición.

Métricas de Calidad

Modelo No Estructurado

Exactitud

Exact_sintactica_bool_json	
Descripción	Evalúa la correctitud sintáctica de un valor
Unidades	{0,1}
Granularidad	Valor

Exact_semantica_bool_json	
Descripción	Evalúa la correctitud semántica de un valor
Unidades	{0,1}
Granularidad	Valor

Exact_precision_json	
Descripción	Evalúa la precisión de un valor
Unidades	{0, 0.25, 0.50, 0.75, 1}
Granularidad	Valor

Completitud

Completitud_cobertura_json	
Descripción	Mide si un archivo cubre todas las entidades que debería cubrir.
Unidades	[0..1]
Granularidad	Archivo

Completitud_densidad_json	
Descripción	Calcula el ratio de valores faltantes de algún atributo con respecto a la cantidad total de documentos. Se consideran todos los documentos, sin importar si el atributo en cuestión está o no definido para todos los documentos.
Unidades	[0..1]
Granularidad	Atributo

Frescura

ratio_actualidad_json	
Descripción	Evalúa la actualidad de un archivo
Unidades	[0..1]
Granularidad	Archivo

Consistencia

Consistencia_Intra_Relacion_json	
Descripción	Evalúa la consistencia entre dos valores distintos
Unidades	{0,1}
Granularidad	Valor

Consistencia_Intra_Relacion_conj_valores_json	
Descripción	Evalúa la consistencia entre un conjunto de valores
Unidades	{0,1}
Granularidad	Conjunto de valores

Unicidad

Unicidad_no_duplicacion_documento_json	
Descripción	Evalúa si un registro es un duplicado exacto o no
Unidades	{0,1}
Granularidad	Documento

Unicidad_no_contradiccion_documento_json	
Descripción	Evalúa si un registro es un duplicado contradictorio o no
Unidades	{0,1}
Granularidad	Documento

Métricas Instanciadas

Modelo No Estructurado

Exactitud

Exact_sintactica_bool_Departamento_receptivos	
Métrica	Exact_sintactica_bool_json
Datos	receptivos.Departamento
Método Medición	if Departamento not in list_departamentos or Departamento != 'Transito' exactitud = 0 else: exactitud = 1

Exact_semántica_bool_`Lugar Ingreso`_receptivos	
Métrica	Exact_semantica_bool_json
Datos	receptivos.`Lugar Ingreso`
Método Medición	if `Lugar Ingreso` not in list_lugares exactitud = 0 else: exactitud = 1

Exact_precision_Pais_receptivos	
Métrica	Exact_precision_json
Datos	receptivos.Pais

Método Medición	if Pais is pais:
	exactitud = 1.00
	elif Pais is Gran Bretaña:
	exactitud = 0.75
	elif "Otro" in Pais:
	exactitud = 0.50
	elif Pais is "Africa u Oceania"
	exactitud = 0.25
	else:
	exactitud = 0

Completitud

Completitud_cobertura_receptivos	
Métrica	Completitud_cobertura_json
Datos	receptivos
Método Medición	cobertura = sum(receptivos['Gente']) / L_ref_estim

Completitud_densidad_Lugar_Ingreso_Receptivos	
Métrica	Completitud_densidad_json
Datos	receptivos."Lugar Ingreso"
Método Medición	densidad = suma_nulos(df["Lugar Ingreso"])

Frescura

ratio_actualidad_receptivos	
Métrica	ratio_actualidad_json
Datos	receptivos
Método Medición	frescura = 1 - (t2-t0) / Δt0

Consistencia

Consistencia_Intra_Relacion_IdFecIng_receptivos	
Métrica	Consistencia_Intra_Relacion_json
Datos	receptivos.'IdFecIng'

Método Medición	if validarDistintosFecIng(IdFecIng) > 1:
	consistencia = 0
	else:
	consistencia = 1

Consistencia_Intra_Relacion_Fechas_Estadia_Receptivos	
Métrica	Consistencia_Intra_Relacion_conj_valores_json
Datos	receptivos.FechaIngreso, receptivos.FechaEgreso, receptivos.Estadia

Método Medición	try:
	if FechaEgreso - FechaIngreso != Estadia
	consistencia = 0
	else:
	consistencia = 1
	except:
	consistencia = 1

Unicidad

Unicidad_no_duplicacion_documento_receptivos	
Métrica	Unicidad_no_duplicacion_documento_json
Datos	receptivos.IdIngresos,receptivos.`Lugar Ingreso`, receptivos.IdTranspIngreso, receptivos.`Transporte Internacional de Ingreso`, receptivos.FechaIngreso, receptivos.IdFecIng,receptivos.FechaEgreso,receptivos.IdFecEgr, receptivos.IdNacionalidad,receptivos.Pais,receptivos.IdResidencia, receptivos.Residencia,receptivos.IdMotivo,receptivos.Motivo, receptivos.IdOcupacion,receptivos.Ocupacion, receptivos.IsEstudio, receptivos.Estudio, receptivos.IdDestinoLocalidad,receptivos.Localidad, receptivos.IdDepartamentoDestino,receptivos.Departamento, receptivos.IdOtroDepartamento,receptivos.`Otro Departamento`, receptivos.IdOtraLocalidad,receptivos.`Otra Localidad`, receptivos.IdAlojamiento,receptivos.Alojamiento, receptivos.IdTranspLocal, receptivos.TransporteLocal, receptivos.IdEgresos, receptivos.`Lugar Egreso` ,receptivos.IdTranspEgreso, receptivos.`Transporte Internacional de Egreso`, receptivos.IdDestino, receptivos.Destino,receptivos.Estadia,receptivos.Gente, receptivos.GastoTotal,receptivos.GastoAlojamiento, receptivos.GastoAlimentacion,receptivos.GastoTransporte,receptivos.GastoCultural,receptivos.GastoTours,receptivos.GastoCompras,receptivos.GastoOtros, receptivos.Coef,receptivos.CoefTot
Método Medición	unicidad = duplicados_row(indice['no_duplicacion']

Unicidad_no_contradiccion_documento_receptivos	
Métrica	Unicidad_no_contradiccion_documento_json
Datos	receptivos.IdIngresos,receptivos.`Lugar Ingreso`, receptivos.IdTranspIngreso, receptivos.`Transporte Internacional de Ingreso`, receptivos.FechaIngreso, receptivos.IdFecIng,receptivos.FechaEgreso,receptivos.IdFecEgr, receptivos.IdNacionalidad,receptivos.Pais,receptivos.IdResidencia, receptivos.Residencia,receptivos.IdMotivo,receptivos.Motivo, receptivos.IdOcupacion,receptivos.Ocupacion, receptivos.IsEstudio, receptivos.Estudio, receptivos.IdDestinoLocalidad,receptivos.Localidad, receptivos.IdDepartamentoDestino,receptivos.Departamento, receptivos.IdOtroDepartamento,receptivos.`Otro Departamento`, receptivos.IdOtraLocalidad,receptivos.`Otra Localidad`, receptivos.IdAlojamiento,receptivos.Alojamiento, receptivos.IdTranspLocal, receptivos.TransporteLocal, receptivos.IdEgresos, receptivos.`Lugar Egreso` ,receptivos.IdTranspEgreso, receptivos.`Transporte Internacional de Egreso`, receptivos.IdDestino, receptivos.Destino,receptivos.Estadia,receptivos.Gente, receptivos.GastoTotal,receptivos.GastoAlojamiento, receptivos.GastoAlimentacion,receptivos.GastoTransporte,receptivos.GastoCultural,receptivos.GastoTours,receptivos.GastoCompras,receptivos.GastoOtros, receptivos.Coef,receptivos.CoefTot
Método Medición	unicidad = duplicados_row(indice)['no_contradiccion']

Métricas Instanciadas

Modelo No Estructurado - Métodos de medición

Exactitud

list_departamentos = ['Montevideo', 'Maldonado', 'Canelones', 'Salto', 'Lavalleja', 'San Jose', 'Rio Negro', nan, 'Durazno', 'Cerro Largo', 'Colonia', 'Tacuarembo', 'Treinta y Tres', 'Rocha', 'Artigas', 'Paysandu', 'Rivera', 'Flores', 'Soriano', 'Florida']

list_lugares = ['Chuy', 'Río Branco', 'Aeropuerto de Carrasco', 'Aeropuerto de Punta del Este', 'Fray Bentos', 'Bella Unión', 'Colonia', 'Salto', 'Paysandú', 'Carmelo', 'Melo - Aceguá', 'Artigas', 'Rivera', 'Puerto de montevideo', 'Nueva Palmira']

Cobertura

Dic 2016	94.131	No hay datos, se estima como dic-21
2017	3.940.790	
2018	3.711.948	
2019	3.220.602	
2020	1.000.908	1er trimestre (hasta comienzo de pandemia)
2021	233.505	Noviembre y diciembre ("fin de pandemia")
2022	2.466.929	
2023	391.683	1er trimestre
Tot	15.060.496	**Valor asignado para L_ref_estim**

Consistencia

```
def validarDistintosFecIng(valorId):
  result = df_receptivos.loc[df_receptivos['IdFecIng'] == valorId]\
           .groupby('IdFecIng')['FechaIngreso'].nunique().reset_index(name='CantDistintos')
  for index,row in result.iterrows():
    if row['CantDistintos'] > 1:
      return 0
    else:
      return 1

def validarDistintosFecEgr(valorId):
  result = df_receptivos.loc[df_receptivos['IdFecEgr'] == valorId]\
           .groupby('IdFecEgr')['FechaEgreso'].nunique().reset_index(name='CantDistintos')
  for index,row in result.iterrows():
    if row['CantDistintos'] > 1:
      return 0
    else:
      return 1
```

Agregaciones de medidas

Modelo Relacional

Dimensión	Nombre	Métrica	Granularidad	Fórmula
Exactitud	ratio_exactitud_sintactica_col	Exact_sintactica_bool	Columna	Suma de valores de medidas de exactitud sintáctica dividido cantidad de valores de la columna
	ratio_precision_col	Exact_precision	Columna	Suma de valores de medidas de precisión dividido cantidad de valores de la columna
	ratio_exactitud_semántica_col	Exact_semantica_bool	Columna	Suma de valores de medidas de exactitud semántica dividido cantidad de valores de la columna
Completitud	ratio_densidad_tabla	Completitud_densidad	Tabla	Promedio de ratios de densidad de las columnas
Consistencia	ratio_consist_columna	Consistencia_Intra_Relacion	Columna	Suma de valores de medidas de consistencia de celdas dividido cantidad de valores de la columna
	ratio_consist_cols_emisivos	Consistencia_Intra_Relacion_conj_celdas	Grupo de columnas	Suma de promedios ponderados de reglas de consistencia intra-relación de cada registro dividido el total de registros. Se establecen los pesos indicados en la fórmula en función de la importancia relativa que se le da a la consistencia entre los distintos atributos.
	ratio_consist_tuplas_receptivos	Consistencia_Intra_Relacion_conj_celdas	Grupo de columnas	Suma de promedios ponderados de reglas de consistencia intra-relación de cada registro dividido el total de registros. Se establecen los pesos indicados en la fórmula en función de la importancia relativa que se le da a la consistencia entre los distintos atributos.
	ratio_int_dom_columna	Consistencia_Int_Dominio	Columna	Suma de valores de medidas de integridad de dominio dividido cantidad de valores de la columna
Unicidad	ratio_no_duplicados	Unicidad_no_duplicacion_tupla	Tabla	Porcentaje de datos que no están duplicados de forma exacta.
	ratio_no_contradicciones	Unicidad_no_contradiccion tupla	Tabla	Porcentaje de datos que no están duplicados con contradicciones

Combinaciones de medidas

Modelo Relacional

Dimensión	Nombre	Métricas	Granularidad	Fórmula
Exactitud	ratio_exactitud	ratio_exact_sintactica_col (s1), ratio_exact_semantica_col (s2)	Columna	α1s1 + α2s2
Unicidad	ratio_unicidad_tabla	ratio_no_duplicados (u1), ratio_no_contradicciones (u2)	Tabla	0.2u1 + 0.8u2

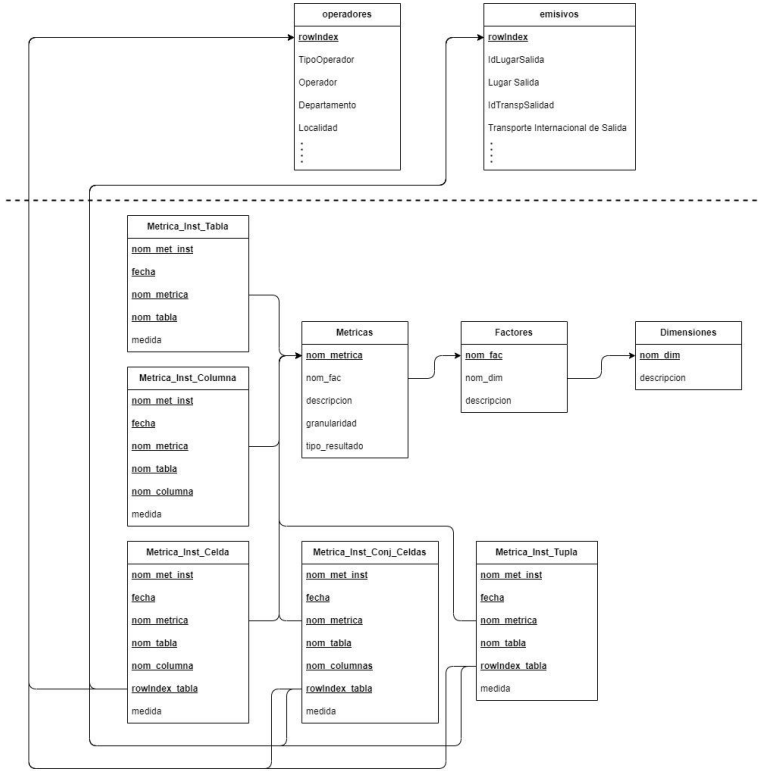
Atributo	Coeficientes	
	α1	α 2
Operadores.Telefono	0.30	0.70
Operadores.Web	0.40	0.60
Operadores.Email	0.40	0.60

Modelo Metadatos Calidad

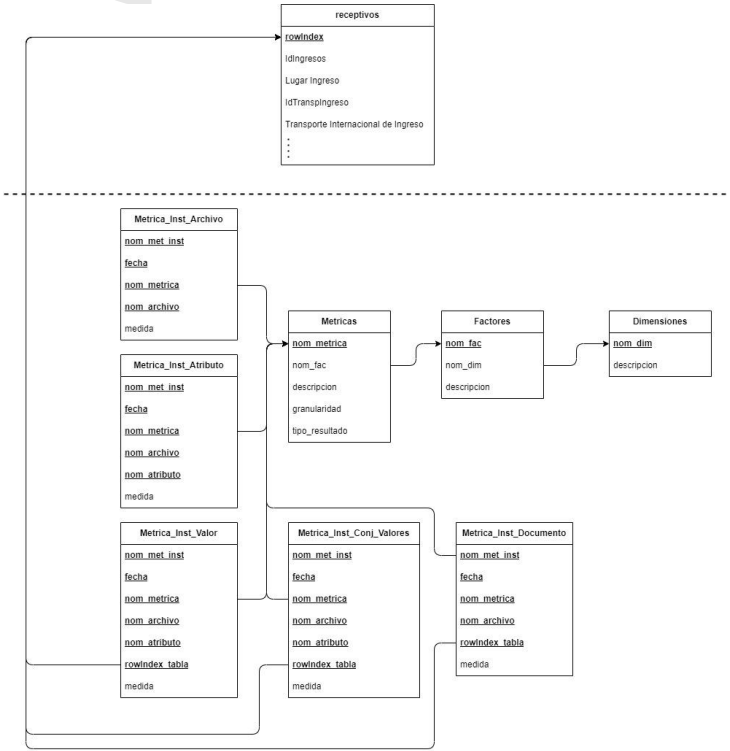
Definición de un modelado de BD que permita el almacenamiento de Dimensiones, Factores, Métricas y Métricas instanciados con los resultados de cada medición, con la finalidad de que esta BD soporte la iteración de las ejecuciones de calidad.

Esto permitirá evaluar la evolución de las mediciones de calidad de nuestros datos.

En este caso dado que nuestros datos de emisivos y operadores no residen en una BD relacional se pensó en el relacionamiento con la BD de Metadatos de Calidad a partir de un atributo rowIndex que es propio de la importación del archivo en la herramienta que se utilizó para analizar los datos, esto debido a que luego de un análisis de los datos no logramos identificar claves candidatas dentro de los archivos.



Modelo Metadatos Calidad



De igual forma que con los archivos anteriores y utilizando el mismo modelo de bd para almacenar las mediciones de calidad, el archivo receptivos, al no encontrarse almacenado en una base relacional, y a diferencia de los dos anteriores siendo este en formato json, el cual se caracteriza por ser flexible en su estructura, aunque este en particular sea bastante similar a emisivos, es que decidimos optar por la misma identificación de los registros a como lo realizamos con los csv.

Calidad de Datos dependiente del Contexto

Dominio: Turismo

Características de Usuario	• U1 - Gerentes • U2 - Administrativos • U3 - Usuarios de los datos

Reglas de Negocio	• RN1 - Todas las fechas deben ser distintas de nulo. • RN2 - Los teléfonos no pueden ser nulos ni vacíos. • RN3 - Los nombres de ciudades y/o departamentos no pueden ser nulos ni vacíos. • RN4 - Los días de estadía deben ser de al menos 1. • RN5 - Los datos registrados no deben contener caracteres especiales. • RN6 - Ninguno de los datos registrados puede estar abreviado. • RN7 - Todos los operadores deben registrar valores de latitud y longitud. • RN8 - Moneda = U$S, es decir, los gastos representan valores en dólares. • RN9 - Si el alojamiento es una vivienda familiar, entonces el costo es 0. • RN10 - La localidad de los operadores determina el departamento. • RN11 - Los estudios de las personas solo pueden ser "primaria", "secundaria" o "terciaria".

Tareas	• T1 - Generación de reportes de los datos • T2 - Registro de datos en la base de datos • T3 - Consultas de los datos

Req. de Sistema	• RS1 - Sistema debe dar resultados de consultas en un tiempo máximo de 2 segundos.

Req. de Calidad de Datos	• RQ1 - Formatos de fechas AAAA/MM/DD • RQ2 - En archivo emisivos Fecha Salida tiene que ser menor a Fecha Entrada. • RQ3 - En archivo receptivos Fecha Ingreso tiene que ser menor a Fecha Egreso • RQ4 - Al menos 10 registros de turismo emisivo por departamento. • RQ5 - Al menos 80% de registros con valores en campo "operador" del archivo operadores turísticos. • RQ6 - Cada operador debe presentar al menos el 50% de sus datos no nulos • RQ7 - Al menos el 90% de los mails ingresados en operadores debe ser válido. • RQ8 - Los gastos de turistas deben ser representados con al menos 2 decimales después de la coma.

Necesidades de Filtrado	• - - -
Metadatos	• M1 - metadatos_parte2.xlsx
Metadatos de Calidad de Datos	• - - -
Otros Datos	• OD1 - agenciasDeViaje.csv • OD2 - agenciasDeViaje-metadatos.xlsx

Calidad de Datos dependiente del Contexto

Cambios o agregados de mediciones en la Especificación del Modelo de Calidad inicial

Primer cambio

Si bien se verifica la dependencia funcional entre las fechas de entrada y salida y la duración de la estadía de los turistas, RN4 resulta en una medición que no tenemos elaborada y deberíamos incluirla para controlar dicho requerimiento. Esta nueva métrica es instanciada a continuación

Métrica instanciada	
Exact_semantica_bool_Estadia_emisivos	
Métrica	Exact_semantica_bool
Datos	emisivos.'Estadia'
Método Medición	if Estadia >= 1 exactitud = 1 else: exactitud = 0

Métrica instanciada	
Exact_semantica_bool_Estadia_receptivos	
Métrica	Exact_semantica_bool
Datos	receptivos.'Estadia'
Método Medición	if Estadia >= 1 exactitud = 1 else: exactitud = 0

Segundo cambio

Debido a lo observado en la etapa de data profiling, originalmente se planteó una métrica de exactitud sintáctica para el campo de FechaSalida del archivo de emisivos y otra para los campos FechaIngreso y FechaEgreso del archivo de receptivos, que verificaban que las fechas estén ingresadas en formato AAAA-MM-DD, ya que éste era el formato de fecha predominante. Sin embargo, se deben realizar dos cambios a estas métricas para poder verificar el cumplimiento del requerimiento RQ1. Estos cambios son los siguientes:

- Modificar la función utilizada en el método de medición de manera de chequear que el formato correcto sea MM/DD/AAAA y no AAAA-MM-DD
- Instanciar la métrica para el campo de FechaEntrada del archivo de emisivos, ya que es el único campo de fecha para el cual no se había instanciado la métrica original (debido a que todos sus valores contaban con el formato considerado como correcto originalmente). A continuación se presentan las métricas instanciadas para los cuatro campos en cuestión y se incluye la función modificada.

Métrica instanciada	
Exact_sintactica_bool_FechaSalida	
Métrica	Exact_sintactica_bool
Datos	emisivos.'FechaSalida'
Método Medición	if check_date_format(FechaSalida) == True exactitud = 1 else: exactitud = 0

Métrica instanciada	
Exact_sintactica_bool_FechaEntrada	
Métrica	Exact_sintactica_bool
Datos	emisivos.'FechaEntrada'
Método Medición	if check_date_format(FechaSalida) == True exactitud = 1 else: exactitud = 0

Tercer cambio

RN5 es un requerimiento que no tuvimos presente, quizás justamente por no haber contado con el contexto suficiente, y creemos que es una de las mediciones que podemos integrar. Para ello, se podría instanciar la métrica de exactitud sintáctica Exact_sintactica_bool con un método de medición alternativo que verifique si el valor contiene caracteres especiales. Si bien esta métrica podría ser instanciada para la gran mayoría de los campos ya tratados, a continuación se ejemplifica esta métrica para uno de los campos para el cual sería aplicable:

Métrica instanciada	
Exact_sintactica_bool_direccion	
Métrica	Exact_sintactica_bool
Datos	operadores.Direccion

Método Medición	if check_no_special_characters(Direccion) == True:
	exactitud = 1
	else
	exactitud = 0

La función utilizada es la siguiente

```
def check_no_special_characters(input_string):
    pattern = r'^[a-zA-Z0-9]+$'
    match = re.match(pattern, str(input_string))
    return match is not None
```

Conclusiones

- El trabajo elaborado estuvo enfocado en tres archivos que contenían información de turismo (turistas que salen e ingresan al país e información de operadores turísticos). Estos archivos contienen información variada en cuanto a sus tipos de datos y cantidad lo cual nos permitió generar varias mediciones de calidad.
- El DataProfiling nos permitió tener una primera aproximación con los datos y entender un poco mas cual era la composición de cada archivo, sobre todo porque no contábamos con mucha información de cuál era el proceso de producción de los datos, tampoco estaban muy claras las potenciales claves que se podían encontrar en cada uno.
- Se generaron estadísticas básicas, y utilizaron algunas técnicas para detectar errores, datos faltantes, duplicados, análisis de posibles claves o si existían claves difusas, todo elaborado en notebooks python.
- Todos estos pasos anteriores nos permitieron entender un poco mas cual era la realidad de los datos a los que necesitábamos medir la calidad incluso diagramar algunas mediciones que eran evidentes que se debían realizar.

Conclusiones

- Para las 5 dimensiones (Exactitud, Completitud, Frescura, Consistencia, Unicidad) estudiadas elaboramos mediciones
  - 10 métricas instanciadas para Exactitud Sintáctica
  - 11 métricas instanciadas para Exactitud Semántica
  - 3 métricas instanciadas para Presición
  - 2 métricas instanciadas para Cobertura
  - 67 métricas instanciadas para Densidad
  - 2 métricas instanciadas para Actualidad
  - 13 métricas instanciadas para Consistencia Intra-Relación
  - 4 métricas instanciadas para Integridad de Dominio
  - 3 métricas instanciadas para No Duplicación
  - 3 métricas instanciadas para No Contradicción
- En la segunda parte del trabajo incluímos al análisis el contexto al cual pertenecen los archivos, esto nos permitió entender un poco más la realidad a la que pertenecen los datos logrando entender por ejemplo el significado de algunos campos a los cuales no sabíamos qué mediciones aplicarle, por ejemplo que la Estadía de un turista no podía ser menor a 1.
- También contar con algunas reglas del negocio y reglas de calidad de datos nos ayudó a enfocar algunas mediciones sobre formatos de datos, o incluso entender algunas relaciones entre algunas columnas cuando adquieren determinados valores para algunos casos que no los tuvimos en cuenta o donde tuvimos que redefinir alguna medición.