Examen calidad de datos María Noel y Carolina

Ejercicio 1 (10 puntos) Una empresa cuenta con una base de datos de clientes potenciales cuyo dato principal de contacto es el email. Dicha empresa tiene como objetivo a corto plazo lanzar una campaña comercial ofreciendo distintos packs de productos a estos clientes potenciales y planifica realizar la misma a través del canal de envío de emails. Se ha detectado que dicho dato fue tomado como un dato de una planilla excel, y al momento de su ingreso en la base de datos (tabla de clientes) no se realizaron controles ni se tuvieron en cuenta patrones de validación. Se le encomienda a su grupo la tarea de analizar la calidad de datos del atributo email en la tabla de clientes y elaborar un plan para mejorar el porcentaje de clientes bien identificados, para lo cual se pide: a) Desarrolle una lista de las principales dimensiones de la calidad de datos vistas en el curso, y los principales factores a tener en cuenta de cada una de ellas. b) ¿Que dimensiones y factores aplicaría en este caso? Justifique. c) Determine métricas y técnicas de medición para poder evaluar el estado en que se encuentra actualmente este dato.

1. Principales dimensiones:

Exactitud: El dato debe refelejar fielmente la realidad que pretende representar. En el caso del email debe tener la dirección correcta.

Completitud: No debe haber datos faltantes, todos los clientes deben tener un email registrado.

Validez: El dato debe cumplir con las reglas del formato predefinidas y tener una estructura valida.

Consistencia: Los datos deben ser consistentes en las diferentes tablas y sistemas un mismo cliente no debería tener diferentes emails en distintas bases de datos.

Actualización: El dato debe estar actualizado y vigente.

Unicidad: No deben existir duplicados.

Accesibilidad: El dato debe ser fácilmente accesible.

1. Dimensiones y factores aplicados.

Validez: Los emails deben tener un formato correcto, se debe verificar que los emails sigan un patrón adecuado para este tipo de registros.

Exactitud: Debemos verificar que el email corresponda a una dirección válida y no contenga errores sintácticos y/o semánticos.

Unicidad: La duplicación de email podría generar el envió repetido a un mismo cliente afectando la campaña.

Actualidad: Si los emails no están actualizados es posible que muchos correos no lleguen a destino afectando así la eficacia de la campaña.

Completitud: Si faltan email la empresa no podría contar con todos los clientes, reduciendo el alcance de la campaña.

Consistencia: Se debe corroborar que no exista más de un registro asociado al mismo email por ejemplo.

1. Métricas y técnicas.

Obtención de nueva información:

Validez: Numero de email que cumplen con el formato correcto.

Se obtiene mediante técnicas de uso de expresión regular para validar el formato del email.

Unicidad: Porcentaje de numero de email duplicado.

Se obtiene mediante el análisis de duplicados en bases de datos.

Actualización: Numero de email que han sido verificados.

Se obtiene mediante el uso de herramientas de verificación de email.

Completitud: Registros completos

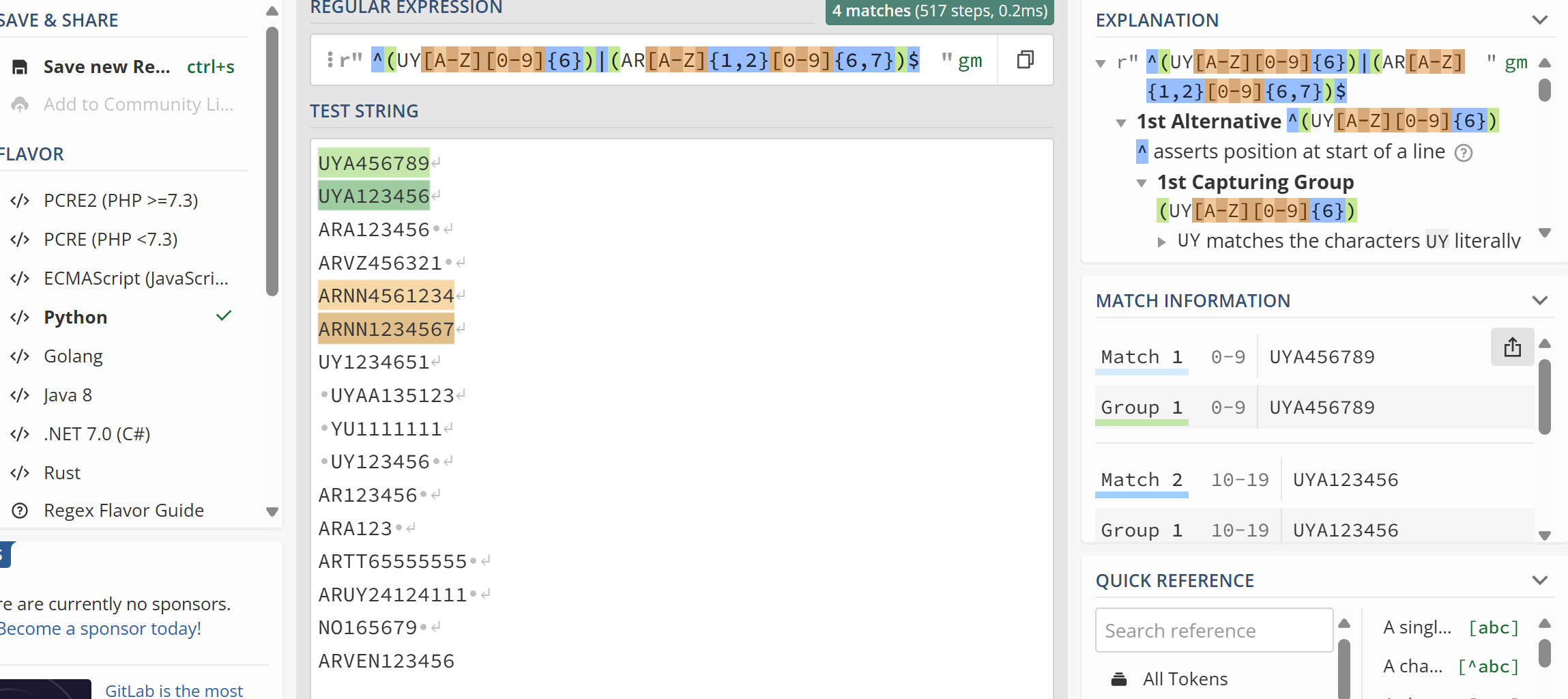
Se obtiene en el análisis de null o valor vacío en la columna de email.

Ejercicio 2 (20 puntos) Desarrolle una expresión regular que valide Pasaportes de Uruguay y Argentina, tomando en cuenta las especificaciones actuales que regulan dicha disposición. Las letras son siempre en mayúsculas.

Uruguay: · Comienza siempre con UY. · Luego de UY siempre sigue una letra más. · A continuación, tienen 6 dígitos numéricos. Entonces seria 3 letras y 6 números siempre Argentina: · Comienza siempre con AR. · luego pueden seguir una o dos letras más. · finaliza con 6 o 7 dígitos numéricos. Entonces seria 3 o 4 letras y 6 o 7 números

Ejemplos Válidos UYA456789 ARA123456 ARVZ456321 ARNN4561234

Ejemplos No Válidos UY1234651 UYAA135123 YU1111111 UY123456 AR123456 ARA123 ARTT65555555 ARUY24124111 NO165679 ARVEN123456



Ejercicio 3 (20 puntos) Dado una columna de datos que se obtienen los teléfonos de varios clientes radicados en varios países, deseamos poder generar una nueva columna donde podamos tener solo el número telefónico sin el código país. Ejemplo: (598)23446777 La idea es crear una columna nueva llamada numero\_telefono donde solo se puede obtener siempre y cuando la columna de datos original contenga estos dos caracteres “(“ y “)”, si no llegase a tener estos dos caracteres debe quedar en blanco la celda de datos.

Ejemplos de cómo quedaría: pone (598)23446777 (59823446777

numero\_telefono 23446777 65454664 (594)11454878 11454878 Sugerencia: usar la función GREL utilizando el if como uno de las funciones más adecuadas, también puede utilizarse otras que ustedes deseen. Conjunto de datos dado (generar proyecto por Clipboard): (55)344555685 (235235456234 2352355 (1)12342378 (1234)1245236 125236 12352350)

Ejercicio 4 (50 puntos) Un municipio cuenta con datos de 2 radares ubicados físicamente en la misma ruta, uno antes del ingreso a la ciudad, Radar 1, donde se permite circular hasta 70 Km/h y el otro, Radar 2, a la salida de la ciudad donde se permite circular hasta 90 KM/h. Dichos radares han detectado infracciones y el municipio las ha recopilado dentro de un solo archivo con la estructura que se muestran en el archivo Radares.csv. Se conoce además que: 1. Los datos que provienen del Radar 1 tienen formato de fecha DD/MM/AAAA mientras que los provenientes del Radar 2 tienen formato de fecha DD-MM-AAAA. 2. Los radares generaron (por motivos desconocidos) algunos registros duplicados, es decir registraron la misma infracción al mismo vehículo en el mismo día y horario más de una vez. 3. Algunas patentes tienen un formato diferente a las esperadas para analizar por el municipio (LL DDDD LL es el formato esperado en este país, donde L es letra y D es digito). Esto puede haber ocurrido por error en el registro del radar o por la circulación de vehículos correspondiente a otros países o bien circulación de vehículos con patentes cuyo formato sea diferente al vigente, razón por la cual estos casos deberán ser analizados de manera individual por personal del municipio. A fines de solucionar estos inconvenientes se solicita: · Unificar los formatos de fechas dejando como formato para todas las transacciones registradas el formato DD/MM/AAAA. · Separar la patente incluida en la columna “Vehiculo e Infraccion” en una columna llamada “Patente”. Considerar que, además del formato esperado (LL DDDD LL), pueden existir registros de casos con menos o más cantidades de letras antes y después de los dígitos y menos o más dígitos entre las letras, pero siempre existirán esos 3 grupos (letras + dígitos + letras) separados por un espacio. · Verificar que no existan valores en blanco dentro de la columna Patente. · Los datos de las patentes que comienzan con SO deben ser eliminados porque los vehículos que tienen patentes con estas características pertenecen a móviles policiales, ambulancias y bomberos. · Eliminar las infracciones duplicadas, dejando solo la primera ocurrencia. Se considera una infracción duplicada si fue realizada por el mismo radar, en la misma fecha y hora al mismo vehículo. · Generar una nueva columna llamada “Patente a analizar” con todos aquellos valores de la columna Patente que no cumplan estrictamente con el formato LL DDDD LL. Para las filas que tengan el formato esperado, el valor en esta columna deberá quedar en blanco (“”). · Extraer de la columna “Vehiculo e Infraccion” la velocidad permitida (solo el numero). · Extraer de la columna “Vehiculo e Infraccion” la velocidad alcanzada (solo el numero). · Generar una columna llamada “Maxima superada en (Km/h)” calculada como: Velocidad Alcanzada – Velocidad Permitida · Eliminar todas las columnas a excepción de las columnas: § Fecha § Hora § Radar § Patente § Patente a analizar § Velocidad permitida § Velocidad alcanzada § Maxima superada en (Km/h) · Exportar los datos al archivo Infracciones.csv con el orden de las columnas según la lista anterior. · Exportar el proyecto