PEC4 - Análisis de Datos en R

Sefora Conti (scontico@uoc.edu) y Salomón Marquez (sblaizer@uoc.edu)

2025-06-19

Contents

Sección 1: Contexto y objetivo del estudio

Describe el ámbito del estudio, el objetivo general, y el conjunto de datos utilizado. Justifica por qué se eligieron esos datos y de dónde provienen. Menciona si hay restricciones de uso o privacidad.

El cáncer gastrointestinal agrupa un conjunto de neoplasias malignas y heterogeneas que afectan distintos órganos del aparato digestivo, incluyendo el tracto biliar, el estómago, el páncreas, el colon y el recto. Estas enfermedades representan una proporción significativa de la carga global del cáncer y presentan, en muchos casos, pronósticos desfavorables (1, 2). La identificación de factores clínico-patológicos que influyen en la supervivencia global de los pacientes es fundamental para mejorar la estratificación de riesgo, la toma de decisiones terapéuticas y el diseño de estrategias de seguimiento personalizadas.

En los últimos años, se han desarrollado diversos enfoques para mejorar esta predicción, incluyendo modelos in vitro (como los organoides derivados de pacientes)(3–5), modelos in vivo (por ejemplo, xenoinjertos en modelos animales)(6) y estrategias in silico(7–9). Estas últimas se apoyan en distintos tipos de datos, tales como imágenes histopatológicas(10), resonancias magnéticas(11) o información clínica, con el fin de generar herramientas que ayuden a personalizar el tratamiento oncológico.

En el presente trabajo se propone analizar un conjunto de datos clínicos disponibles públicamente a través del repositorio de TCGA (The Cancer Genome Atlas Program), con el objetivo de identificar posibles factores clínicos asociados a la supervivencia de pacientes con tumores gastrointestinales. Este análisis explorará posibles correlaciones entre variables patológicas y demográficas sobre la supervivencia global de los pacientes, con el fin de aportar evidencia que permita mejorar la estratificación de pacientes.

Los datos utilizados son de acceso libre, están anonimizados y su uso con fines académicos está permitido según las políticas del GDC Data Portal.

Bibliografía:

1. M. F. Bijlsma, A. Sadanandam, P. Tan, L. Vermeulen, Molecular subtypes in cancers of the gastrointestinal tract. Nat Rev Gastroenterol Hepatol 14, 333–342 (2017).
2. R. L. Siegel, K. D. Miller, H. E. Fuchs, A. Jemal, Cancer statistics, 2022. CA: A Cancer Journal for Clinicians 72, 7–33 (2022).
3. J. Drost, H. Clevers, Organoids in cancer research. Nat Rev Cancer 18, 407–418 (2018).
4. P. W. Nagle, J. Th. M. Plukker, C. T. Muijs, P. van Luijk, R. P. Coppes, Patient-derived tumor organoids for prediction of cancer treatment response. Seminars in Cancer Biology 53, 258–264 (2018).
5. C. A. Pasch, P. F. Favreau, A. E. Yueh, C. P. Babiarz, A. A. Gillette, J. T. Sharick, M. R. Karim, K. P. Nickel, A. K. DeZeeuw, C. M. Sprackling, P. B. Emmerich, R. A. DeStefanis, R. T. Pitera, S. N. Payne, D. P. Korkos, L. Clipson, C. M. Walsh, D. Miller, E. H. Carchman, M. E. Burkard, K. K. Lemmon, K. A. Matkowskyj, M. A. Newton, I. M. Ong, M. F. Bassetti, R. J. Kimple, M. C. Skala, D. A. Deming, Patient-Derived Cancer Organoid Cultures to Predict Sensitivity to Chemotherapy and Radiation. Clinical Cancer Research 25, 5376–5387 (2019).
6. E. R. Zanella, E. Grassi, L. Trusolino, Towards precision oncology with patient-derived xenografts. Nat Rev Clin Oncol 19, 719–732 (2022).
7. X. Sun, B. Hu, Mathematical modeling and computational prediction of cancer drug resistance. Briefings in Bioinformatics 19, 1382–1399 (2018).
8. E. J. Mucaki, J. Z. L. Zhao, D. J. Lizotte, P. K. Rogan, Predicting responses to platin chemotherapy agents with biochemically-inspired machine learning. Sig Transduct Target Ther 4, 1–12 (2019).
9. C. Yang, X. Huang, Y. Li, J. Chen, Y. Lv, S. Dai, Prognosis and personalized treatment prediction in TP53-mutant hepatocellular carcinoma: an in silico strategy towards precision oncology. Briefings in Bioinformatics 22, bbaa164 (2021).
10. F. Li, Y. Yang, Y. Wei, P. He, J. Chen, Z. Zheng, H. Bu, Deep learning-based predictive biomarker of pathological complete response to neoadjuvant chemotherapy from histological images in breast cancer. J Transl Med 19, 348 (2021). 11. Z. Liu, Z. Li, J. Qu, R. Zhang, X. Zhou, L. Li, K. Sun, Z. Tang, H. Jiang, H. Li, Q. Xiong, Y. Ding, X. Zhao, K. Wang, Z. Liu, J. Tian, Radiomics of Multiparametric MRI for Pretreatment Prediction of Pathologic Complete Response to Neoadjuvant Chemotherapy in Breast Cancer: A Multicenter Study. Clinical Cancer Research 25, 3538–3547 (2019).

Sección 2: Prospección y preparación de los datos

2.1 Exploración inicial del dataset

La siguiente tabla muestra una descripción general del dataset TCGA. En las siguientes secciones se muestra cómo se obtuvo esta información.

Puntos clave	Descripción
Tipo de archivo y datos	El conjunto de datos fue importado desde un archivo .csv delimitado por punto y coma, que contiene información clínica y patológica de pacientes oncológicos. Incluye variables como edad, tipo de cáncer, presencia de invasión venosa o perineural, inestabilidad de microsatélites, y variables de seguimiento como días hasta el último contacto o la muerte. Debido a su origen combinado (diferentes fuentes clínicas dentro de TCGA), el conjunto de datos requiere un preprocesamiento considerable para hacerlo apto para el análisis.
Variables disponibles	Variables categóricas (por ejemplo, tipo de cáncer, sexo) y variables continuas (edad, peso, altura, etc.).
Dimensiones del dataset	El conjunto de datos incluye 154 variables clínicas de 1.308 pacientes de las cuáles hemos elegido 34 variables que consideramos son las más relevantes para el objetivo de nuestro estudio.
Detección de valores nulos	Hay 135.291 valores nulos en el conjunto de datos original y 21.007 en el nuevo conjunto de datos (con las 34 variables seleccionadas).
Inconsistencias	Se detectaron inconsistencias típicas en datos clínicos, como discordancias entre el estado vital y los días de seguimiento o muerte. Para poder calcular la supervivencia global, fue necesario crear una nueva variable de tiempo de supervivencia (`os_time`) combinando `days_to_death` y `days_to_last_followup`, así como una variable binaria de evento (`os_event`).

```r
# Cargar el conjunto de datos (csv)
df_TCGA <- read.csv("Combined_TCGA_Clinical_300525.csv",header=TRUE, sep = ";")
# Visualizar registros
head(df_TCGA[,1:5], n=3) #mostramos los primeros 3 registros
```

```
##   bcr_patient_barcode additional_studies tumor_tissue_site
## 1        TCGA-3L-AA1B               <NA>             Colon
## 2        TCGA-4N-A93T               <NA>             Colon
## 3        TCGA-4T-AA8H               <NA>             Colon
##              histological_type other_dx
## 1          Colon Adenocarcinoma       No
## 2          Colon Adenocarcinoma       No
## 3 Colon Mucinous Adenocarcinoma       No
```

```r
tail(df_TCGA[,1:5], n=3) #mostramos los ultimos 3 registros
```

```
##      bcr_patient_barcode additional_studies tumor_tissue_site
## 1306        TCGA-Z5-AAPL               <NA>          Pancreas
## 1307                <NA>               <NA>              <NA>
## 1308
##                         histological_type other_dx
## 1306 Pancreas-Adenocarcinoma Ductal Type       No
## 1307                                <NA>     <NA>
## 1308
```

```r
# Obtener el número de variables y sus nombres
dim(df_TCGA) #dimensiones de df
```

```
## [1] 1308  154
```

```
names(df_TCGA) #nombres de las variables
```

```
##   [1] "bcr_patient_barcode"
##   [2] "additional_studies"
##   [3] "tumor_tissue_site"
##   [4] "histological_type"
##   [5] "other_dx"
##   [6] "gender"
##   [7] "vital_status"
##   [8] "days_to_birth"
##   [9] "days_to_last_known_alive"
##  [10] "days_to_death"
##  [11] "days_to_last_followup"
##  [12] "race_list"
##  [13] "tissue_source_site"
##  [14] "patient_id"
##  [15] "bcr_patient_uuid"
##  [16] "history_of_neoadjuvant_treatment"
##  [17] "informed_consent_verified"
##  [18] "icd_o_3_site"
##  [19] "icd_o_3_histology"
##  [20] "icd_10"
##  [21] "tissue_prospective_collection_indicator"
##  [22] "tissue_retrospective_collection_indicator"
##  [23] "days_to_initial_pathologic_diagnosis"
##  [24] "age_at_initial_pathologic_diagnosis"
##  [25] "year_of_initial_pathologic_diagnosis"
##  [26] "person_neoplasm_cancer_status"
##  [27] "ethnicity"
##  [28] "weight"
##  [29] "height"
##  [30] "day_of_form_completion"
##  [31] "month_of_form_completion"
##  [32] "year_of_form_completion"
##  [33] "residual_tumor"
##  [34] "anatomic_neoplasm_subdivision"
##  [35] "primary_lymph_node_presentation_assessment"
##  [36] "lymph_node_examined_count"
##  [37] "number_of_lymphnodes_positive_by_he"
##  [38] "number_of_lymphnodes_positive_by_ihc"
##  [39] "preoperative_pretreatment_cea_level"
##  [40] "non_nodal_tumor_deposits"
##  [41] "circumferential_resection_margin"
##  [42] "venous_invasion"
##  [43] "lymphatic_invasion"
##  [44] "perineural_invasion_present"
##  [45] "microsatellite_instability"
##  [46] "number_of_loci_tested"
##  [47] "number_of_abnormal_loci"
##  [48] "kras_gene_analysis_performed"
##  [49] "kras_mutation_found"
##  [50] "kras_mutation_codon"
```

```
##  [51] "braf_gene_analysis_performed"
##  [52] "braf_gene_analysis_result"
##  [53] "synchronous_colon_cancer_present"
##  [54] "history_of_colon_polyps"
##  [55] "colon_polyps_present"
##  [56] "loss_expression_of_mismatch_repair_proteins_by_ihc"
##  [57] "loss_expression_of_mismatch_repair_proteins_by_ihc_results"
##  [58] "number_of_first_degree_relatives_with_cancer_diagnosis"
##  [59] "radiation_therapy"
##  [60] "postoperative_rx_tx"
##  [61] "primary_therapy_outcome_success"
##  [62] "has_new_tumor_events_information"
##  [63] "has_drugs_information"
##  [64] "has_radiations_information"
##  [65] "has_follow_ups_information"
##  [66] "project"
##  [67] "stage_event_system_version"
##  [68] "stage_event_clinical_stage"
##  [69] "stage_event_pathologic_stage"
##  [70] "stage_event_tnm_categories"
##  [71] "stage_event_psa"
##  [72] "stage_event_gleason_grading"
##  [73] "stage_event_ann_arbor"
##  [74] "stage_event_serum_markers"
##  [75] "stage_event_igcccg_stage"
##  [76] "stage_event_masaoka_stage"
##  [77] "cancer_type"
##  [78] "patient_death_reason"
##  [79] "anatomic_neoplasm_subdivision_other"
##  [80] "neoplasm_histologic_grade"
##  [81] "country_of_procurement"
##  [82] "city_of_procurement"
##  [83] "reflux_history"
##  [84] "antireflux_treatment"
##  [85] "antireflux_treatment_types"
##  [86] "barretts_esophagus"
##  [87] "h_pylori_infection"
##  [88] "family_history_of_stomach_cancer"
##  [89] "number_of_relatives_with_stomach_cancer"
##  [90] "relative_family_cancer_history"
##  [91] "cancer_first_degree_relative"
##  [92] "blood_relative_cancer_history_list"
##  [93] "history_hepato_carcinoma_risk_factors"
##  [94] "post_op_ablation_embolization_tx"
##  [95] "eastern_cancer_oncology_group"
##  [96] "primary_pathology_tumor_tissue_site"
##  [97] "primary_pathology_histological_type"
##  [98] "primary_pathology_specimen_collection_method_name"
##  [99] "primary_pathology_history_prior_surgery_type_other"
## [100] "primary_pathology_days_to_initial_pathologic_diagnosis"
## [101] "primary_pathology_age_at_initial_pathologic_diagnosis"
## [102] "primary_pathology_year_of_initial_pathologic_diagnosis"
## [103] "primary_pathology_neoplasm_histologic_grade"
## [104] "primary_pathology_residual_tumor"
```

```
## [105] "primary_pathology_vascular_tumor_cell_type"
## [106] "primary_pathology_perineural_invasion_present"
## [107] "primary_pathology_child_pugh_classification_grade"
## [108] "primary_pathology_ca_19_9_level"
## [109] "primary_pathology_ca_19_9_level_lower"
## [110] "primary_pathology_ca_19_9_level_upper"
## [111] "primary_pathology_fetoprotein_outcome_value"
## [112] "primary_pathology_fetoprotein_outcome_lower_limit"
## [113] "primary_pathology_fetoprotein_outcome_upper_limit"
## [114] "primary_pathology_platelet_result_count"
## [115] "primary_pathology_platelet_result_lower_limit"
## [116] "primary_pathology_platelet_result_upper_limit"
## [117] "primary_pathology_prothrombin_time_result_value"
## [118] "primary_pathology_inter_norm_ratio_lower_limit"
## [119] "primary_pathology_intern_norm_ratio_upper_limit"
## [120] "primary_pathology_albumin_result_specified_value"
## [121] "primary_pathology_albumin_result_lower_limit"
## [122] "primary_pathology_albumin_result_upper_limit"
## [123] "primary_pathology_bilirubin_upper_limit"
## [124] "primary_pathology_bilirubin_lower_limit"
## [125] "primary_pathology_total_bilirubin_upper_limit"
## [126] "primary_pathology_creatinine_value_in_mg_dl"
## [127] "primary_pathology_creatinine_lower_level"
## [128] "primary_pathology_creatinine_upper_limit"
## [129] "primary_pathology_fibrosis_ishak_score"
## [130] "primary_pathology_cholangitis_tissue_evidence"
## [131] "adenocarcinoma_invasion"
## [132] "histological_type_other"
## [133] "tumor_type"
## [134] "initial_pathologic_diagnosis_method"
## [135] "init_pathology_dx_method_other"
## [136] "surgery_performed_type"
## [137] "histologic_grading_tier_category"
## [138] "maximum_tumor_dimension"
## [139] "source_of_patient_death_reason"
## [140] "tobacco_smoking_history"
## [141] "year_of_tobacco_smoking_onset"
## [142] "stopped_smoking_year"
## [143] "number_pack_years_smoked"
## [144] "alcohol_history_documented"
## [145] "alcoholic_exposure_category"
## [146] "frequency_of_alcohol_consumption"
## [147] "amount_of_alcohol_consumption_per_day"
## [148] "history_of_diabetes"
## [149] "days_to_diabetes_onset"
## [150] "history_of_chronic_pancreatitis"
## [151] "days_to_pancreatitis_onset"
## [152] "family_history_of_cancer"
## [153] "relative_cancer_types"
## [154] "history_prior_surgery_type_other"
```

```r
# Visualizar estructura del conjunto de datos y un resumen estadístico
str(df_TCGA)
```

```
## 'data.frame':    1308 obs. of  154 variables:
##  $ bcr_patient_barcode                                : chr  "TCGA-3L-AA1B" "TCGA-4N-A93T" "TC
##  $ additional_studies                                 : chr  NA NA NA NA ...
##  $ tumor_tissue_site                                  : chr  "Colon" "Colon" "Colon" "Colon"
##  $ histological_type                                  : chr  "Colon Adenocarcinoma" "Colon Ad
##  $ other_dx                                           : chr  "No" "No" "No" "No" ...
##  $ gender                                             : chr  "FEMALE" "MALE" "FEMALE" "MALE"
##  $ vital_status                                       : chr  NA NA NA "Dead" ...
##  $ days_to_birth                                      : int  -22379 -24523 -15494 -27095 -148
##  $ days_to_last_known_alive                           : int  NA NA NA NA NA NA 424 NA NA NA .
##  $ days_to_death                                      : int  NA NA NA 49 290 NA NA 1126 NA NA
##  $ days_to_last_followup                              : num  475 146 385 -Inf -Inf ...
##  $ race_list                                          : chr  "BLACK OR AFRICAN AMERICAN" "BLA
##  $ tissue_source_site                                 : chr  "3L" "4N" "4T" "5M" ...
##  $ patient_id                                         : chr  "AA1B" "A93T" "AA8H" "AAT4" ...
##  $ bcr_patient_uuid                                   : chr  "A94E1279-A975-480A-93E9-7B1FF05
##  $ history_of_neoadjuvant_treatment                   : chr  "No" "No" "No" "No" ...
##  $ informed_consent_verified                          : chr  "YES" "YES" "YES" "YES" ...
##  $ icd_o_3_site                                       : chr  "C18.0" "C18.2" "C18.6" "C18.2"
##  $ icd_o_3_histology                                  : chr  "01/03/8140" "01/03/8140" "01/03
##  $ icd_10                                             : chr  "C18.0" "C18.2" "C18.6" "C18.2"
##  $ tissue_prospective_collection_indicator            : chr  "YES" "YES" "NO" "NO" ...
##  $ tissue_retrospective_collection_indicator          : chr  "NO" "NO" "YES" "YES" ...
##  $ days_to_initial_pathologic_diagnosis               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age_at_initial_pathologic_diagnosis                : int  61 67 42 74 40 76 45 85 82 71 ..
##  $ year_of_initial_pathologic_diagnosis               : int  2013 2013 2013 2009 2009 2011 20
##  $ person_neoplasm_cancer_status                      : chr  "TUMOR FREE" "WITH TUMOR" "TUMOR
##  $ ethnicity                                          : chr  "NOT HISPANIC OR LATINO" "NOT HI
##  $ weight                                             : num  63.3 134 108 NA 99.1 ...
##  $ height                                             : num  173 168 168 NA 162 ...
##  $ day_of_form_completion                             : int  22 1 5 27 27 27 14 28 4 15 ...
##  $ month_of_form_completion                           : int  4 10 6 1 1 1 10 1 10 10 ...
##  $ year_of_form_completion                            : int  2014 2014 2014 2015 2015 2015 20
##  $ residual_tumor                                     : chr  "R0" "R0" "R0" "R0" ...
##  $ anatomic_neoplasm_subdivision                      : chr  "Cecum" "Ascending Colon" "Desce
##  $ primary_lymph_node_presentation_assessment         : chr  "YES" "YES" "YES" "YES" ...
##  $ lymph_node_examined_count                          : int  28 25 24 3 11 15 22 27 29 20 ...
##  $ number_of_lymphnodes_positive_by_he                : int  0 NA 0 0 10 0 NA 3 1 7 ...
##  $ number_of_lymphnodes_positive_by_ihc               : int  0 2 NA 0 0 0 NA NA NA 0 ...
##  $ preoperative_pretreatment_cea_level                : num  NA 2 NA 550 2.61 2.91 NA 17.4 3.
##  $ non_nodal_tumor_deposits                           : chr  "NO" "YES" "NO" "NO" ...
##  $ circumferential_resection_margin                   : num  NA 30 20 NA NA NA NA NA NA NA ..
##  $ venous_invasion                                    : chr  "NO" "NO" "NO" "YES" ...
##  $ lymphatic_invasion                                 : chr  "NO" "NO" "NO" NA ...
##  $ perineural_invasion_present                        : chr  "NO" "NO" "NO" NA ...
##  $ microsatellite_instability                         : chr  "NO" NA "NO" NA ...
##  $ number_of_loci_tested                              : int  NA NA NA NA NA NA NA NA NA NA ..
##  $ number_of_abnormal_loci                            : int  NA NA NA NA NA NA NA NA NA NA ..
##  $ kras_gene_analysis_performed                       : chr  "NO" "NO" "NO" "NO" ...
##  $ kras_mutation_found                                : chr  NA NA NA NA ...
##  $ kras_mutation_codon                                : int  NA NA NA NA NA NA NA NA NA NA ..
##  $ braf_gene_analysis_performed                       : chr  "NO" "NO" "NO" "NO" ...
##  $ braf_gene_analysis_result                          : chr  NA NA NA NA ...
##  $ synchronous_colon_cancer_present                   : chr  "NO" "YES" "NO" "NO" ...
```

```
##  $ history_of_colon_polyps                                    : chr  "YES" "NO" "NO" "NO" ...
##  $ colon_polyps_present                                       : chr  "YES" "YES" "NO" "YES" ...
##  $ loss_expression_of_mismatch_repair_proteins_by_ihc         : chr  "YES" "YES" "YES" "NO" ...
##  $ loss_expression_of_mismatch_repair_proteins_by_ihc_results: chr  "MLH1-ExpressedMSH2-ExpressedPMS
##  $ number_of_first_degree_relatives_with_cancer_diagnosis     : int  0 0 0 0 NA 0 0 0 0 0 ...
##  $ radiation_therapy                                          : chr  "NO" "NO" "NO" "NO" ...
##  $ postoperative_rx_tx                                        : chr  "NO" "YES" "NO" "NO" ...
##  $ primary_therapy_outcome_success                            : chr  "Complete Remission/Response" "S
##  $ has_new_tumor_events_information                           : chr  "NO" "NO" "NO" "NO" ...
##  $ has_drugs_information                                      : chr  "NO" "YES" "NO" "NO" ...
##  $ has_radiations_information                                 : chr  "NO" "NO" "NO" "NO" ...
##  $ has_follow_ups_information                                 : chr  "YES" "YES" "YES" "YES" ...
##  $ project                                                    : chr  "TCGA-COAD" "TCGA-COAD" "TCGA-CO
##  $ stage_event_system_version                                 : chr  "7th" "7th" "7th" "6th" ...
##  $ stage_event_clinical_stage                                 : logi  NA NA NA NA NA NA ...
##  $ stage_event_pathologic_stage                               : chr  "Stage I" "Stage IIIB" "Stage II
##  $ stage_event_tnm_categories                                 : chr  "T2N0M0" "T4aN1bM0" "T3N0MX" "T3
##  $ stage_event_psa                                            : logi  NA NA NA NA NA NA ...
##  $ stage_event_gleason_grading                                : logi  NA NA NA NA NA NA ...
##  $ stage_event_ann_arbor                                      : logi  NA NA NA NA NA NA ...
##  $ stage_event_serum_markers                                  : logi  NA NA NA NA NA NA ...
##  $ stage_event_igcccg_stage                                   : logi  NA NA NA NA NA NA ...
##  $ stage_event_masaoka_stage                                  : logi  NA NA NA NA NA NA ...
##  $ cancer_type                                                : chr  "TCGA-COAD" "TCGA-COAD" "TCGA-CO
##  $ patient_death_reason                                       : chr  NA NA NA NA ...
##  $ anatomic_neoplasm_subdivision_other                        : chr  NA NA NA NA ...
##  $ neoplasm_histologic_grade                                  : chr  NA NA NA NA ...
##  $ country_of_procurement                                     : chr  NA NA NA NA ...
##  $ city_of_procurement                                        : chr  NA NA NA NA ...
##  $ reflux_history                                             : chr  NA NA NA NA ...
##  $ antireflux_treatment                                       : chr  NA NA NA NA ...
##  $ antireflux_treatment_types                                 : chr  NA NA NA NA ...
##  $ barretts_esophagus                                         : chr  NA NA NA NA ...
##  $ h_pylori_infection                                         : chr  NA NA NA NA ...
##  $ family_history_of_stomach_cancer                           : chr  NA NA NA NA ...
##  $ number_of_relatives_with_stomach_cancer                    : int  NA NA NA NA NA NA NA NA NA NA ..
##  $ relative_family_cancer_history                             : chr  NA NA NA NA ...
##  $ cancer_first_degree_relative                               : int  NA NA NA NA NA NA NA NA NA NA ..
##  $ blood_relative_cancer_history_list                         : chr  NA NA NA NA ...
##  $ history_hepato_carcinoma_risk_factors                      : chr  NA NA NA NA ...
##  $ post_op_ablation_embolization_tx                           : chr  NA NA NA NA ...
##  $ eastern_cancer_oncology_group                              : int  NA NA NA NA NA NA NA NA NA NA ..
##  $ primary_pathology_tumor_tissue_site                        : chr  NA NA NA NA ...
##  $ primary_pathology_histological_type                        : chr  NA NA NA NA ...
##  $ primary_pathology_specimen_collection_method_name          : chr  NA NA NA NA ...
##  $ primary_pathology_history_prior_surgery_type_other         : chr  NA NA NA NA ...
##   [list output truncated]
```

```
summary(df_TCGA)
```

```
##  bcr_patient_barcode additional_studies tumor_tissue_site  histological_type
##  Length:1308         Length:1308        Length:1308        Length:1308
##  Class :character    Class :character   Class :character   Class :character
##  Mode  :character    Mode  :character   Mode  :character   Mode  :character
```

```
##
##
##
##
##    other_dx            gender          vital_status       days_to_birth
##  Length:1308        Length:1308        Length:1308        Min.   :-32873
##  Class :character   Class :character   Class :character   1st Qu.:-27321
##  Mode  :character   Mode  :character   Mode  :character   Median :-24655
##                                                           Mean   :-24204
##                                                           3rd Qu.:-21319
##                                                           Max.   :-10659
##                                                           NA's   :17
##  days_to_last_known_alive days_to_death    days_to_last_followup
##  Min.   :   0.0           Min.   :   0.0   Min.   :-Inf
##  1st Qu.: 343.5           1st Qu.: 145.0   1st Qu.:   0
##  Median : 992.0           Median : 366.0   Median : 485
##  Mean   :1635.1           Mean   : 503.5   Mean   :-Inf
##  3rd Qu.:3223.0           3rd Qu.: 641.0   3rd Qu.: 942
##  Max.   :3920.0           Max.   :3042.0   Max.   :4502
##  NA's   :1289             NA's   :1071     NA's   :205
##   race_list         tissue_source_site  patient_id        bcr_patient_uuid
##  Length:1308        Length:1308        Length:1308        Length:1308
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  history_of_neoadjuvant_treatment informed_consent_verified icd_o_3_site
##  Length:1308                      Length:1308               Length:1308
##  Class :character                 Class :character          Class :character
##  Mode  :character                 Mode  :character          Mode  :character
##
##
##
##
##  icd_o_3_histology     icd_10          tissue_prospective_collection_indicator
##  Length:1308        Length:1308        Length:1308
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  tissue_retrospective_collection_indicator days_to_initial_pathologic_diagnosis
##  Length:1308                                Min.   :0
##  Class :character                           1st Qu.:0
##  Mode  :character                           Median :0
##                                             Mean   :0
##                                             3rd Qu.:0
##                                             Max.   :0
##                                             NA's   :58
##  age_at_initial_pathologic_diagnosis year_of_initial_pathologic_diagnosis
##  Min.   :30.00                       Min.   :1996
```

```
##  1st Qu.:58.00                       1st Qu.:2008
##  Median :67.00                       Median :2010
##  Mean   :65.85                       Mean   :2009
##  3rd Qu.:74.00                       3rd Qu.:2011
##  Max.   :90.00                       Max.   :2013
##  NA's   :56                          NA's   :53
##  person_neoplasm_cancer_status  ethnicity            weight
##  Length:1308                    Length:1308        Min.   : 34.00
##  Class :character               Class :character   1st Qu.: 65.00
##  Mode  :character               Mode  :character   Median : 78.05
##                                                    Mean   : 80.44
##                                                    3rd Qu.: 91.85
##                                                    Max.   :175.30
##                                                    NA's   :936
##      height      day_of_form_completion month_of_form_completion
##  Min.   : 80.3   Min.   : 1.00          Min.   : 1.000
##  1st Qu.:162.0   1st Qu.:11.00          1st Qu.: 4.000
##  Median :170.0   Median :18.00          Median : 6.000
##  Mean   :168.9   Mean   :16.62          Mean   : 5.881
##  3rd Qu.:176.0   3rd Qu.:22.00          3rd Qu.: 7.000
##  Max.   :193.0   Max.   :31.00          Max.   :12.000
##  NA's   :957     NA's   :41             NA's   :5
##  year_of_form_completion residual_tumor     anatomic_neoplasm_subdivision
##  Min.   :2010            Length:1308        Length:1308
##  1st Qu.:2011            Class :character   Class :character
##  Median :2011            Mode  :character   Mode  :character
##  Mean   :2012
##  3rd Qu.:2013
##  Max.   :2015
##  NA's   :5
##  primary_lymph_node_presentation_assessment lymph_node_examined_count
##  Length:1308                                Min.   :  0.00
##  Class :character                           1st Qu.: 12.00
##  Mode  :character                           Median : 18.00
##                                             Mean   : 21.42
##                                             3rd Qu.: 27.00
##                                             Max.   :109.00
##                                             NA's   :138
##  number_of_lymphnodes_positive_by_he number_of_lymphnodes_positive_by_ihc
##  Min.   : 0.000                      Min.   : 0.0000
##  1st Qu.: 0.000                      1st Qu.: 0.0000
##  Median : 1.000                      Median : 0.0000
##  Mean   : 3.493                      Mean   : 0.2833
##  3rd Qu.: 4.000                      3rd Qu.: 0.0000
##  Max.   :57.000                      Max.   :12.0000
##  NA's   :143                         NA's   :1188
##  preoperative_pretreatment_cea_level non_nodal_tumor_deposits
##  Min.   :   0.000                    Length:1308
##  1st Qu.:   1.700                    Class :character
##  Median :   3.160                    Mode  :character
##  Mean   :  65.074
##  3rd Qu.:   8.982
##  Max.   :7868.000
##  NA's   :906
```

```
##  circumferential_resection_margin venous_invasion    lymphatic_invasion
##  Min.   :  0.00                   Length:1308        Length:1308
##  1st Qu.:  2.50                   Class :character   Class :character
##  Median : 13.00                   Mode  :character   Mode  :character
##  Mean   : 22.96
##  3rd Qu.: 30.00
##  Max.   :165.00
##  NA's   :1185
##  perineural_invasion_present microsatellite_instability number_of_loci_tested
##  Length:1308                 Length:1308                Min.   : 0.000
##  Class :character            Class :character           1st Qu.: 5.000
##  Mode  :character            Mode  :character           Median : 5.000
##                                                         Mean   : 4.944
##                                                         3rd Qu.: 5.000
##                                                         Max.   :10.000
##                                                         NA's   :1236
##  number_of_abnormal_loci kras_gene_analysis_performed kras_mutation_found
##  Min.   :0.0000          Length:1308                  Length:1308
##  1st Qu.:0.0000          Class :character             Class :character
##  Median :0.0000          Mode  :character             Mode  :character
##  Mean   :0.5915
##  3rd Qu.:0.0000
##  Max.   :9.0000
##  NA's   :1237
##  kras_mutation_codon braf_gene_analysis_performed braf_gene_analysis_result
##  Min.   :12.00       Length:1308                  Length:1308
##  1st Qu.:12.00       Class :character             Class :character
##  Median :12.00       Mode  :character             Mode  :character
##  Mean   :13.83
##  3rd Qu.:12.00
##  Max.   :61.00
##  NA's   :1278
##  synchronous_colon_cancer_present history_of_colon_polyps colon_polyps_present
##  Length:1308                      Length:1308             Length:1308
##  Class :character                 Class :character        Class :character
##  Mode  :character                 Mode  :character        Mode  :character
##
##
##
##
##  loss_expression_of_mismatch_repair_proteins_by_ihc
##  Length:1308
##  Class :character
##  Mode  :character
##
##
##
##
##  loss_expression_of_mismatch_repair_proteins_by_ihc_results
##  Length:1308
##  Class :character
##  Mode  :character
##
##
```

```
##
##
##  number_of_first_degree_relatives_with_cancer_diagnosis radiation_therapy
##  Min.   :0.0000                                         Length:1308
##  1st Qu.:0.0000                                         Class :character
##  Median :0.0000                                         Mode  :character
##  Mean   :0.1654
##  3rd Qu.:0.0000
##  Max.   :3.0000
##  NA's   :770
##  postoperative_rx_tx primary_therapy_outcome_success
##  Length:1308         Length:1308
##  Class :character    Class :character
##  Mode  :character    Mode  :character
##
##
##
##
##  has_new_tumor_events_information has_drugs_information
##  Length:1308                      Length:1308
##  Class :character                 Class :character
##  Mode  :character                 Mode  :character
##
##
##
##
##  has_radiations_information has_follow_ups_information   project
##  Length:1308                Length:1308                Length:1308
##  Class :character           Class :character           Class :character
##  Mode  :character           Mode  :character           Mode  :character
##
##
##
##
##  stage_event_system_version stage_event_clinical_stage
##  Length:1308                Mode:logical
##  Class :character           NA's:1308
##  Mode  :character
##
##
##
##
##  stage_event_pathologic_stage stage_event_tnm_categories stage_event_psa
##  Length:1308                  Length:1308                Mode:logical
##  Class :character             Class :character           NA's:1308
##  Mode  :character             Mode  :character
##
##
##
##
##  stage_event_gleason_grading stage_event_ann_arbor stage_event_serum_markers
##  Mode:logical                Mode:logical          Mode:logical
##  NA's:1308                   NA's:1308             NA's:1308
##
```

```
##
##
##
##
##  stage_event_igcccg_stage stage_event_masaoka_stage cancer_type
##  Mode:logical             Mode:logical              Length:1308
##  NA's:1308                NA's:1308                 Class :character
##                                                     Mode  :character
##
##
##
##
##  patient_death_reason anatomic_neoplasm_subdivision_other
##  Length:1308          Length:1308
##  Class :character     Class :character
##  Mode  :character     Mode  :character
##
##
##
##
##  neoplasm_histologic_grade country_of_procurement city_of_procurement
##  Length:1308               Length:1308            Length:1308
##  Class :character          Class :character       Class :character
##  Mode  :character          Mode  :character       Mode  :character
##
##
##
##
##  reflux_history     antireflux_treatment antireflux_treatment_types
##  Length:1308        Length:1308          Length:1308
##  Class :character   Class :character     Class :character
##  Mode  :character   Mode  :character     Mode  :character
##
##
##
##
##  barretts_esophagus h_pylori_infection family_history_of_stomach_cancer
##  Length:1308        Length:1308        Length:1308
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  number_of_relatives_with_stomach_cancer relative_family_cancer_history
##  Min.   :0.0000                          Length:1308
##  1st Qu.:0.0000                          Class :character
##  Median :0.0000                          Mode  :character
##  Mean   :0.2632
##  3rd Qu.:0.0000
##  Max.   :2.0000
##  NA's   :1232
##  cancer_first_degree_relative blood_relative_cancer_history_list
##  Min.   :1.000                Length:1308
```

```
##  1st Qu.:1.000                Class :character
##  Median :1.000                Mode  :character
##  Mean   :1.696
##  3rd Qu.:2.000
##  Max.   :4.000
##  NA's   :1285
##  history_hepato_carcinoma_risk_factors post_op_ablation_embolization_tx
##  Length:1308                           Length:1308
##  Class :character                      Class :character
##  Mode  :character                      Mode  :character
##
##
##
##
##  eastern_cancer_oncology_group primary_pathology_tumor_tissue_site
##  Min.   :0.0000                Length:1308
##  1st Qu.:0.0000                Class :character
##  Median :0.0000                Mode  :character
##  Mean   :0.3514
##  3rd Qu.:1.0000
##  Max.   :3.0000
##  NA's   :1271
##  primary_pathology_histological_type
##  Length:1308
##  Class :character
##  Mode  :character
##
##
##
##
##  primary_pathology_specimen_collection_method_name
##  Length:1308
##  Class :character
##  Mode  :character
##
##
##
##
##  primary_pathology_history_prior_surgery_type_other
##  Length:1308
##  Class :character
##  Mode  :character
##
##
##
##
##  primary_pathology_days_to_initial_pathologic_diagnosis
##  Min.   :0
##  1st Qu.:0
##  Median :0
##  Mean   :0
##  3rd Qu.:0
##  Max.   :0
##  NA's   :1260
```

```
##  primary_pathology_age_at_initial_pathologic_diagnosis
##  Min.   :29.00
##  1st Qu.:58.00
##  Median :66.00
##  Mean   :63.64
##  3rd Qu.:73.00
##  Max.   :82.00
##  NA's   :1263
##  primary_pathology_year_of_initial_pathologic_diagnosis
##  Min.   :2005
##  1st Qu.:2010
##  Median :2011
##  Mean   :2011
##  3rd Qu.:2012
##  Max.   :2013
##  NA's   :1260
##  primary_pathology_neoplasm_histologic_grade primary_pathology_residual_tumor
##  Length:1308                                 Length:1308
##  Class :character                            Class :character
##  Mode  :character                            Mode  :character
##
##
##
##
##  primary_pathology_vascular_tumor_cell_type
##  Length:1308
##  Class :character
##  Mode  :character
##
##
##
##
##  primary_pathology_perineural_invasion_present
##  Length:1308
##  Class :character
##  Mode  :character
##
##
##
##
##  primary_pathology_child_pugh_classification_grade
##  Length:1308
##  Class :character
##  Mode  :character
##
##
##
##
##  primary_pathology_ca_19_9_level primary_pathology_ca_19_9_level_lower
##  Min.   :   1.0                  Min.   :0.0000
##  1st Qu.:  26.5                  1st Qu.:0.0000
##  Median :  54.9                  Median :0.0000
##  Mean   : 345.8                  Mean   :0.1351
##  3rd Qu.: 200.0                  3rd Qu.:0.0000
```

```
##  Max.   :6910.0                  Max.   :5.0000
##  NA's   :1268                    NA's   :1271
##  primary_pathology_ca_19_9_level_upper
##  Min.   : 7.00
##  1st Qu.:35.00
##  Median :55.00
##  Mean   :45.56
##  3rd Qu.:55.00
##  Max.   :55.00
##  NA's   :1267
##  primary_pathology_fetoprotein_outcome_value
##  Min.   : 1.380
##  1st Qu.: 2.350
##  Median : 3.100
##  Mean   : 3.803
##  3rd Qu.: 4.325
##  Max.   :14.000
##  NA's   :1280
##  primary_pathology_fetoprotein_outcome_lower_limit
##  Min.   :0
##  1st Qu.:0
##  Median :0
##  Mean   :0
##  3rd Qu.:0
##  Max.   :0
##  NA's   :1275
##  primary_pathology_fetoprotein_outcome_upper_limit
##  Min.   : 6.000
##  1st Qu.: 6.000
##  Median : 6.000
##  Mean   : 7.579
##  3rd Qu.: 9.000
##  Max.   :15.000
##  NA's   :1275
##  primary_pathology_platelet_result_count
##  Min.   :   134.0
##  1st Qu.:   211.0
##  Median :   269.5
##  Mean   : 68013.6
##  3rd Qu.:179500.0
##  Max.   :354000.0
##  NA's   :1264
##  primary_pathology_platelet_result_lower_limit
##  Min.   :   140
##  1st Qu.:   150
##  Median :   150
##  Mean   : 32560
##  3rd Qu.: 15000
##  Max.   :150000
##  NA's   :1263
##  primary_pathology_platelet_result_upper_limit
##  Min.   :   400
##  1st Qu.:   450
##  Median :   450
```

```
##  Mean   :126091
##  3rd Qu.:400000
##  Max.   :450000
##  NA's   :1263
##  primary_pathology_prothrombin_time_result_value
##  Min.   : 0.800
##  1st Qu.: 1.000
##  Median : 1.100
##  Mean   : 2.012
##  3rd Qu.: 1.100
##  Max.   :12.200
##  NA's   :1266
##  primary_pathology_inter_norm_ratio_lower_limit
##  Min.   : 0.000
##  1st Qu.: 0.800
##  Median : 0.900
##  Mean   : 1.865
##  3rd Qu.: 0.900
##  Max.   :10.400
##  NA's   :1274
##  primary_pathology_intern_norm_ratio_upper_limit
##  Min.   : 1.100
##  1st Qu.: 1.200
##  Median : 1.200
##  Mean   : 2.553
##  3rd Qu.: 1.200
##  Max.   :13.100
##  NA's   :1274
##  primary_pathology_albumin_result_specified_value
##  Min.   :2.400
##  1st Qu.:3.775
##  Median :4.150
##  Mean   :3.998
##  3rd Qu.:4.325
##  Max.   :4.800
##  NA's   :1268
##  primary_pathology_albumin_result_lower_limit
##  Min.   :3.300
##  1st Qu.:3.500
##  Median :3.500
##  Mean   :3.467
##  3rd Qu.:3.500
##  Max.   :3.500
##  NA's   :1268
##  primary_pathology_albumin_result_upper_limit
##  Min.   :4.50
##  1st Qu.:4.80
##  Median :5.00
##  Mean   :4.91
##  3rd Qu.:5.00
##  Max.   :5.20
##  NA's   :1268
##  primary_pathology_bilirubin_upper_limit
##  Min.   : 0.200
```

```
##  1st Qu.: 0.400
##  Median : 0.700
##  Mean   : 2.859
##  3rd Qu.: 1.025
##  Max.   :84.000
##  NA's   :1264
##  primary_pathology_bilirubin_lower_limit
##  Min.   : 0.000
##  1st Qu.: 0.100
##  Median : 0.100
##  Mean   : 2.221
##  3rd Qu.: 0.150
##  Max.   :78.000
##  NA's   :1265
##  primary_pathology_total_bilirubin_upper_limit
##  Min.   : 0.300
##  1st Qu.: 1.000
##  Median : 1.000
##  Mean   : 5.505
##  3rd Qu.: 1.200
##  Max.   :96.000
##  NA's   :1265
##  primary_pathology_creatinine_value_in_mg_dl
##  Min.   :0.5000
##  1st Qu.:0.8000
##  Median :0.9000
##  Mean   :0.8651
##  3rd Qu.:1.0000
##  Max.   :1.4000
##  NA's   :1265
##  primary_pathology_creatinine_lower_level
##  Min.   :0.4000
##  1st Qu.:0.6000
##  Median :0.6000
##  Mean   :0.6349
##  3rd Qu.:0.7000
##  Max.   :0.9000
##  NA's   :1265
##  primary_pathology_creatinine_upper_limit
##  Min.   :1.000
##  1st Qu.:1.100
##  Median :1.100
##  Mean   :1.198
##  3rd Qu.:1.300
##  Max.   :1.400
##  NA's   :1265
##  primary_pathology_fibrosis_ishak_score
##  Length:1308
##  Class :character
##  Mode  :character
##
##
##
##
```

```
##  primary_pathology_cholangitis_tissue_evidence adenocarcinoma_invasion
##  Length:1308                                   Length:1308
##  Class :character                              Class :character
##  Mode  :character                              Mode  :character
##
##
##
##
##  histological_type_other  tumor_type        initial_pathologic_diagnosis_method
##  Length:1308              Length:1308       Length:1308
##  Class :character         Class :character  Class :character
##  Mode  :character         Mode  :character  Mode  :character
##
##
##
##
##  init_pathology_dx_method_other surgery_performed_type
##  Length:1308                    Length:1308
##  Class :character               Class :character
##  Mode  :character               Mode  :character
##
##
##
##
##  histologic_grading_tier_category maximum_tumor_dimension
##  Length:1308                        Min.   : 0.300
##  Class :character                   1st Qu.: 2.925
##  Mode  :character                   Median : 3.500
##                                     Mean   : 3.840
##                                     3rd Qu.: 4.500
##                                     Max.   :14.000
##                                     NA's   :1138
##  source_of_patient_death_reason tobacco_smoking_history
##  Length:1308                    Min.   :1.000
##  Class :character               1st Qu.:1.000
##  Mode  :character               Median :2.000
##                                 Mean   :2.201
##                                 3rd Qu.:3.000
##                                 Max.   :5.000
##                                 NA's   :1159
##  year_of_tobacco_smoking_onset stopped_smoking_year number_pack_years_smoked
##  Min.   :1948                  Min.   :1952         Min.   : 0.30
##  1st Qu.:1960                  1st Qu.:1980         1st Qu.:15.00
##  Median :1971                  Median :1988         Median :25.00
##  Mean   :1971                  Mean   :1990         Mean   :26.84
##  3rd Qu.:1982                  3rd Qu.:2004         3rd Qu.:40.00
##  Max.   :1993                  Max.   :2013         Max.   :75.00
##  NA's   :1261                  NA's   :1258         NA's   :1251
##  alcohol_history_documented alcoholic_exposure_category
##  Length:1308                Length:1308
##  Class :character           Class :character
##  Mode  :character           Mode  :character
##
##
```

```
##
##
##  frequency_of_alcohol_consumption amount_of_alcohol_consumption_per_day
##  Min.   :0.500                    Min.   :0.500
##  1st Qu.:3.000                    1st Qu.:1.000
##  Median :6.500                    Median :1.000
##  Mean   :4.812                    Mean   :1.581
##  3rd Qu.:7.000                    3rd Qu.:2.000
##  Max.   :7.000                    Max.   :4.000
##  NA's   :1276                     NA's   :1277
##  history_of_diabetes days_to_diabetes_onset history_of_chronic_pancreatitis
##  Length:1308         Min.   :-9070.00       Length:1308
##  Class :character    1st Qu.: -163.00       Class :character
##  Mode  :character    Median :  -52.00       Mode  :character
##                      Mean   : -706.79
##                      3rd Qu.:  -17.75
##                      Max.   :  504.00
##                      NA's   :1294
##  days_to_pancreatitis_onset family_history_of_cancer relative_cancer_types
##  Min.   :-18029.0           Length:1308              Length:1308
##  1st Qu.:  -231.5           Class :character         Class :character
##  Median :   -71.0           Mode  :character         Mode  :character
##  Mean   : -1744.2
##  3rd Qu.:   -38.0
##  Max.   :     1.0
##  NA's   :1297
##  history_prior_surgery_type_other
##  Length:1308
##  Class :character
##  Mode  :character
##
##
##
##
```

```
# Verificar valores nulos
print(table(is.na(df_TCGA)))
```

```
##
##  FALSE   TRUE
##  66141 135291
```

```
# Verificar valores nulos por columnas
# print(colSums(is.na(df_TCGA)))

# Verificar valores nulos por filas
# print(rowSums(is.na(df_TCGA)))
```

Identificamos la columna que sirva con ID principal de cada registro del dataset

```
# Visualizar los primeros 5 elementos de las siguientes columnas tipo ID
head(df_TCGA[,c("bcr_patient_barcode", "bcr_patient_uuid", "patient_id")], 5)
```

```
##   bcr_patient_barcode                  bcr_patient_uuid patient_id
## 1        TCGA-3L-AA1B A94E1279-A975-480A-93E9-7B1FF05CBCBF       AA1B
## 2        TCGA-4N-A93T 92554413-9EBC-4354-8E1B-9682F3A031D9       A93T
## 3        TCGA-4T-AA8H A5E14ADD-1552-4606-9FFE-3A03BCF76640       AA8H
## 4        TCGA-5M-AAT4 1136DD50-242A-4659-AAD4-C53F9E759BB3       AAT4
## 5        TCGA-5M-AAT6 CE00896A-F7D2-4123-BB95-24CB6E53FC32       AAT6
```

En general observamos tres diferentes identificadores: el TCGA barcode completo, UUID (Universal Unique Identifier) y código interno de paciente, respectivamente. En este proyecto utilizaremos el primero de ellos pues es el identificador estándar a nivel de paciente en TCGA. Además observamos que también contiene el código interno de paciente.

```
# Crear un subconjunto del dataset TCGA con variables relevantes desde una perspectiva oncologica
df_gastro <-subset(df_TCGA, select = c(bcr_patient_barcode, tumor_tissue_site, histological_type,
                                gender, vital_status, days_to_death, days_to_last_followup,
                                tissue_source_site,
                                age_at_initial_pathologic_diagnosis, person_neoplasm_cancer_status,
                                weight, height, residual_tumor, anatomic_neoplasm_subdivision,
                                number_of_lymphnodes_positive_by_he, number_of_lymphnodes_positive_by_ih
                                preoperative_pretreatment_cea_level, non_nodal_tumor_deposits,
                                circumferential_resection_margin, venous_invasion, lymphatic_invasion,
                                perineural_invasion_present, microsatellite_instability,
                                number_of_loci_tested, number_of_abnormal_loci, kras_gene_analysis_perfo
                                kras_mutation_found, kras_mutation_codon, braf_gene_analysis_performed,
                                braf_gene_analysis_result, synchronous_colon_cancer_present,
                                stage_event_pathologic_stage, stage_event_tnm_categories, cancer_type))
```

Visualizamos con mayor detalle el TCGA dataset personalizado con las 34 variables de interés.

```
# Usar función skimr()
skim(df_gastro)
```

```
## Warning: There was 1 warning in `dplyr::summarize()`.
## i In argument: `dplyr::across(tidyselect::any_of(variable_names),
##   mangled_skimmers$funs)`.
## i In group 0: .
## Caused by warning:
## ! There was 1 warning in `dplyr::summarize()`.
## i In argument: `dplyr::across(tidyselect::any_of(variable_names),
##   mangled_skimmers$funs)`.
## Caused by warning in `inline_hist()`:
## ! Variable contains Inf or -Inf value(s) that were converted to NA.
```

Table 2: Data summary

Name	df_gastro
Number of rows	1308
Number of columns	34
Column type frequency:	
character	22

numeric	12
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
bcr_patient_barcode	1	1.00	0	12	1	1307	0
tumor_tissue_site	6	1.00	0	9	1	6	0
histological_type	17	0.99	0	65	1	19	0
gender	2	1.00	0	6	1	3	0
vital_status	636	0.51	4	5	0	4	0
tissue_source_site	1	1.00	0	2	1	99	0
person_neoplasm_cancer_status	194	0.85	0	10	1	3	0
residual_tumor	142	0.89	0	2	1	5	0
anatomic_neoplasm_subdivision	81	0.94	0	25	1	19	0
non_nodal_tumor_deposits	1010	0.23	0	3	1	3	0
venous_invasion	762	0.42	0	3	1	3	0
lymphatic_invasion	740	0.43	0	3	1	3	0
perineural_invasion_present	1030	0.21	0	3	1	3	0
microsatellite_instability	1190	0.09	0	3	1	3	0
kras_gene_analysis_performed	733	0.44	0	3	1	3	0
kras_mutation_found	1245	0.05	0	3	1	3	0
braf_gene_analysis_performed	747	0.43	0	3	1	3	0
braf_gene_analysis_result	1272	0.03	0	8	1	3	0
synchronous_colon_cancer_present	744	0.43	0	3	1	3	0
stage_event_pathologic_stage	52	0.96	0	10	1	15	0
stage_event_tnm_categories	3	1.00	0	9	1	141	0
cancer_type	1	1.00	0	9	1	6	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
days_to_death	1071	0.18	503.46	524.96	0.0	145.0	366.00	641.00	3042.0	
days_to_last_followup	205	0.84	-Inf	NaN	-Inf	0.0	485.00	942.00	4502.0	
age_at_initial_pathologic_diagnosis	56	0.96	65.85	11.88	30.0	58.0	67.00	74.00	90.0	
weight	936	0.28	80.44	20.88	34.0	65.0	78.05	91.85	175.3	
height	957	0.27	168.87	11.64	80.3	162.0	170.00	176.00	193.0	
number_of_lymphnodes_positive_by_he	143	0.89	3.49	6.26	0.0	0.0	1.00	4.00	57.0	
number_of_lymphnodes_positive_by_ihc	1188	0.09	0.28	1.64	0.0	0.0	0.00	0.00	12.0	
preoperative_pretreatment_cea_level	906	0.31	65.07	468.84	0.0	1.7	3.16	8.98	7868.0	
circumferential_resection_margin	1185	0.09	22.96	28.31	0.0	2.5	13.00	30.00	165.0	
number_of_loci_tested	1236	0.06	4.94	2.14	0.0	5.0	5.00	5.00	10.0	
number_of_abnormal_loci	1237	0.05	0.59	1.70	0.0	0.0	0.00	0.00	9.0	
kras_mutation_codon	1278	0.02	13.83	8.92	12.0	12.0	12.00	12.00	61.0	

```
# Obtener el número de variables y sus nombres
dim(df_gastro)
```

```
## [1] 1308   34
```

```
# Verificar valores nulos
print(table(is.na(df_gastro)))
```

```
##
## FALSE  TRUE
## 23465 21007
```

Transformación de los datos

La primera transformación al dataframe `df_gastro` consiste en la creación de una nueva variable de supervivencia de los pacientes `os_time` para identificar los que siguen vivos y los que no. Veamos,

```
# Visualizar algunas variables relacionadas a la supervivencia de pacientes
df_gastro[1:10, c("vital_status", "days_to_last_followup", "days_to_death")]
```

```
##    vital_status days_to_last_followup days_to_death
## 1          <NA>                  475            NA
## 2          <NA>                  146            NA
## 3          <NA>                  385            NA
## 4          Dead                 -Inf            49
## 5          Dead                 -Inf           290
## 6          <NA>                 1200            NA
## 7          <NA>                  775            NA
## 8          Dead                 1126          1126
## 9          <NA>                 1419            NA
## 10         <NA>                 1331            NA
```

donde,

- **vital_status**: estado de vida del paciente al momento del cierre del seguimiento
- **days_to_last_followup**: número de días desde el diagnóstico hasta la última vez que se tuvo contacto clínico con el paciente
- **days_to_death**: número de días desde el diagnóstico hasta la muerte del paciente

La variable `os_time` es una combinación de las variables `days_to_last_followup` y `days_to_death` donde se imputan los valores `NA` de `days_to_death` por aquellos en `days_to_last_followup`.

```
# Crear la variable de supervivencia de pacientes (los que siguen vivos y los que no)
df_gastro$os_time <- ifelse(
  !is.na(df_gastro$days_to_death), #si hay un valor que no es nulo en la variable "days_to_death"
  df_gastro$days_to_death,
  df_gastro$days_to_last_followup #en caso contrario, el valor es lo que aparece en la columna "days_to_
)

# Transformar los valores infinitos de os_time en NA
df_gastro$os_time[is.infinite(df_gastro$os_time)] <- NA
```

En los datos clínicos de TCGA, en ocasiones la variable `vital_status` no se informa o se analiza incorrectamente, pero si los datos de seguimiento del paciente están presentes, se puede asumir que el paciente estaba vivo en su último seguimiento. Este es el supuesto tomado en cuenta en este proyecto, aunque en un enfoque más conservador, la estrategia sería descartar a estos pacientes.

```r
# Asignar "Alive" a los NAs de vital_status si la variable "days_to_last_followup" tiene un valor valid
df_gastro$vital_status[is.na(df_gastro$vital_status)
                       & !is.na(df_gastro$days_to_last_followup)] <- "Alive"

# Definir os_event, si el paciente está vivo con 0, caso contrario con 1.
df_gastro$os_event <- ifelse(df_gastro$vital_status == "Dead", 1, 0)

# Eliminar la especificación TCGA del tipo de cáncer para simplificar.
df_gastro$cancer_type <-gsub("TCGA-","", df_gastro$cancer_type) #ref. The R book page 124

# Eliminar la especificación Stage para simplificar.
df_gastro$stage_event_pathologic_stage <-gsub("Stage ","", df_gastro$stage_event_pathologic_stage) #ref

# Estructura del conjunto de datos y resumen estadístico
str(df_gastro)
```

```
## 'data.frame':    1308 obs. of  36 variables:
##  $ bcr_patient_barcode                : chr  "TCGA-3L-AA1B" "TCGA-4N-A93T" "TCGA-4T-AA8H" "TCGA-5M-
##  $ tumor_tissue_site                  : chr  "Colon" "Colon" "Colon" "Colon" ...
##  $ histological_type                  : chr  "Colon Adenocarcinoma" "Colon Adenocarcinoma" "Colon Mu
##  $ gender                             : chr  "FEMALE" "MALE" "FEMALE" "MALE" ...
##  $ vital_status                       : chr  "Alive" "Alive" "Alive" "Dead" ...
##  $ days_to_death                      : int  NA NA NA 49 290 NA NA 1126 NA NA ...
##  $ days_to_last_followup              : num  475 146 385 -Inf -Inf ...
##  $ tissue_source_site                 : chr  "3L" "4N" "4T" "5M" ...
##  $ age_at_initial_pathologic_diagnosis : int  61 67 42 74 40 76 45 85 82 71 ...
##  $ person_neoplasm_cancer_status      : chr  "TUMOR FREE" "WITH TUMOR" "TUMOR FREE" "WITH TUMOR" ..
##  $ weight                             : num  63.3 134 108 NA 99.1 ...
##  $ height                             : num  173 168 168 NA 162 ...
##  $ residual_tumor                     : chr  "R0" "R0" "R0" "R0" ...
##  $ anatomic_neoplasm_subdivision      : chr  "Cecum" "Ascending Colon" "Descending Colon" "Ascending
##  $ number_of_lymphnodes_positive_by_he : int  0 NA 0 0 10 0 NA 3 1 7 ...
##  $ number_of_lymphnodes_positive_by_ihc: int  0 2 NA 0 0 0 NA NA NA 0 ...
##  $ preoperative_pretreatment_cea_level : num  NA 2 NA 550 2.61 2.91 NA 17.4 3.4 32.8 ...
##  $ non_nodal_tumor_deposits           : chr  "NO" "YES" "NO" "NO" ...
##  $ circumferential_resection_margin   : num  NA 30 20 NA NA NA NA NA NA NA ...
##  $ venous_invasion                    : chr  "NO" "NO" "NO" "YES" ...
##  $ lymphatic_invasion                 : chr  "NO" "NO" "NO" NA ...
##  $ perineural_invasion_present        : chr  "NO" "NO" "NO" NA ...
##  $ microsatellite_instability         : chr  "NO" NA "NO" NA ...
##  $ number_of_loci_tested              : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ number_of_abnormal_loci            : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ kras_gene_analysis_performed       : chr  "NO" "NO" "NO" "NO" ...
##  $ kras_mutation_found                : chr  NA NA NA NA ...
##  $ kras_mutation_codon                : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ braf_gene_analysis_performed       : chr  "NO" "NO" "NO" "NO" ...
##  $ braf_gene_analysis_result          : chr  NA NA NA NA ...
##  $ synchronous_colon_cancer_present   : chr  "NO" "YES" "NO" "NO" ...
##  $ stage_event_pathologic_stage       : chr  "I" "IIIB" "IIA" "IV" ...
##  $ stage_event_tnm_categories         : chr  "T2N0M0" "T4aN1bM0" "T3N0MX" "T3N0M1b" ...
##  $ cancer_type                        : chr  "COAD" "COAD" "COAD" "COAD" ...
##  $ os_time                            : num  475 146 385 49 290 ...
##  $ os_event                           : num  0 0 0 1 1 0 0 1 0 0 ...
```

2.2 Preguntas objetivo

1. ¿Qué correlación existe entre el estadio patológico del tumor en el momento del diagnóstico y la supervivencia?
2. ¿Qué diferencias en supervivencia se observan según el género y la edad de los pacientes?
3. ¿Existen diferencias significativas en la supervivencia según el tipo de cáncer gastrointestinal?
4. ¿Cuál es el impacto de la invasión venosa, linfática o perineural sobre la supervivencia?

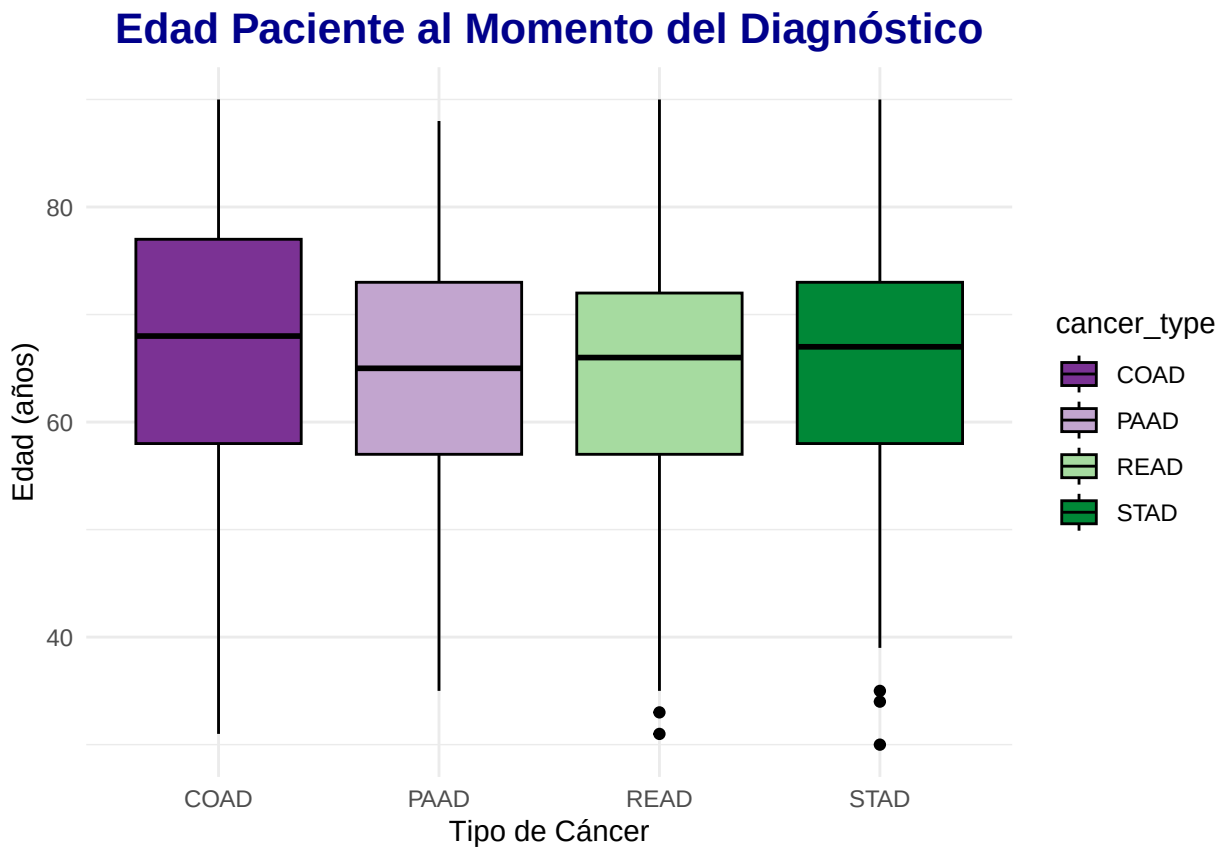
Sección 3: Análisis exploratorio de los datos

3.1 Estadística descriptiva y visualización

En esta sección implementamos un análisis exploratorio y de visualización de datos con el objetivo de responder a las preguntas objetivo de la sección 2.2.

```
# Convertir a factor las variables tipo `chr`
df_gastro$gender<- factor(df_gastro$gender)
df_gastro$residual_tumor<- factor(df_gastro$residual_tumor)
df_gastro$venous_invasion<- factor(df_gastro$venous_invasion)
df_gastro$lymphatic_invasion<- factor(df_gastro$lymphatic_invasion)
df_gastro$perineural_invasion_present<- factor(df_gastro$perineural_invasion_present)
df_gastro$microsatellite_instability<- factor(df_gastro$microsatellite_instability)
df_gastro$kras_mutation_found<- factor(df_gastro$kras_mutation_found)
df_gastro$stage_event_pathologic_stage<- factor(df_gastro$stage_event_pathologic_stage)
df_gastro$cancer_type<- factor(df_gastro$cancer_type)
```

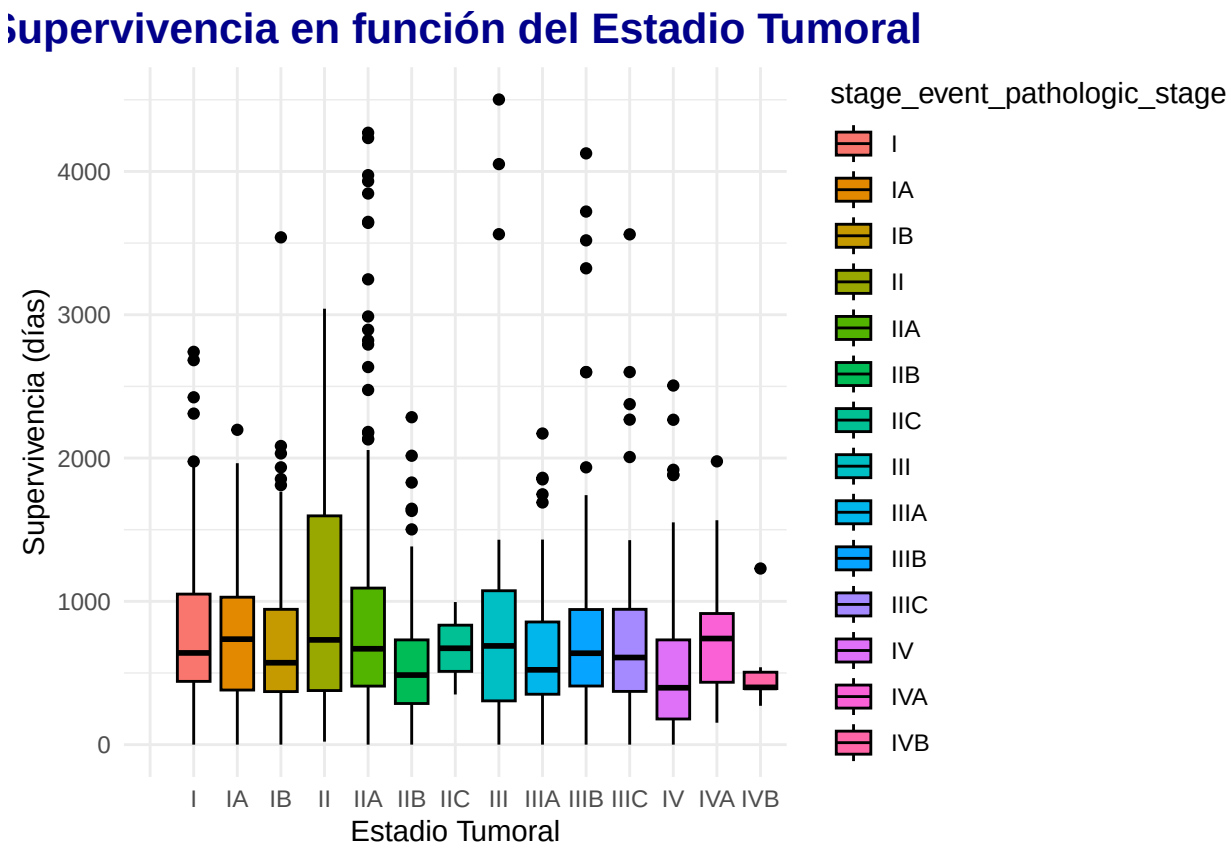
```
# Visualizar la distribucion de edades en los distintos tipos de cancer
ggplot(data=subset(df_gastro, !is.na(age_at_initial_pathologic_diagnosis)), aes(x = cancer_type, y = ag
geom_boxplot(col = 'black')+
  labs(title = "Edad Paciente al Momento del Diagnóstico", x ="Tipo de Cáncer", y = "Edad (años)") +
  theme_minimal()+
  scale_fill_brewer(palette="PRGn")+
  theme(plot.title = element_text(size=16, color='Darkblue', face='bold', hjust = 0.5))
```



Pregunta 1: ¿Qué correlación existe entre el estadio patológico del tumor en el momento del diagnóstico y la supervivencia?

```
# Correlacion entre el estadio tumoral y la supervivencia de los pacientes
ggplot(data=subset(df_gastro, !is.na(stage_event_pathologic_stage)), aes(x = stage_event_pathologic_stag
geom_boxplot(col = 'black')+
  labs(title = "Supervivencia en función del Estadio Tumoral", x ="Estadio Tumoral", y = "Supervivencia
  theme_minimal()+
  #scale_fill_brewer(palette="PRGn")+
  theme(plot.title = element_text(size=16, color='Darkblue', face='bold', hjust = 0.5))
```

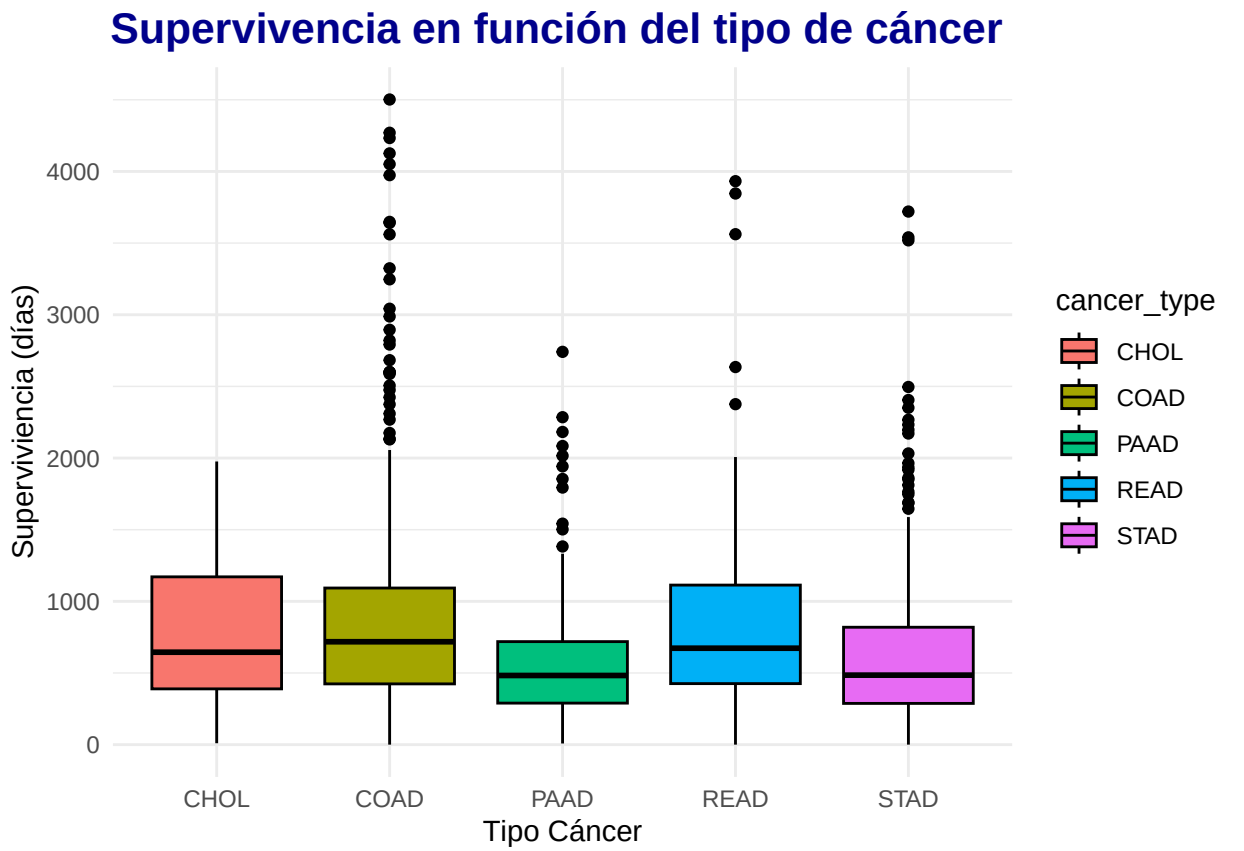
```
## Warning: Removed 225 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



El gráfico de supervivencia en función de estadio tumoral muestra que la supervivencia disminuye progresivamente a medida que aumenta el estadio tumoral, tal como lo señalan Singh et. al (https://doi.org/10.5114/pg.2024.141834). De aquí la importancia crítica de realizar un diagnóstico precoz en pacientes oncológicos. Para aportar mayor respaldo estadístico a esta observación, sería necesario realizar un análisis más riguroso usando un modelo de riesgos proporcionales de Cox. Dicho análisis queda fuera del alcance de este proyecto.

Pregunta 2: ¿Existen diferencias significativas en la supervivencia según el tipo de cáncer gastrointestinal?

```
# Supervivencia en función del tipo de cáncer
ggplot(data=subset(df_gastro, !is.na(os_time)), aes(x = cancer_type, y = os_time, fill = cancer_type))+
geom_boxplot(col = 'black')+
  labs(title = "Supervivencia en función del tipo de cáncer ", x ="Tipo Cáncer", y = "Superviviencia (d
  theme_minimal()+
  #scale_fill_brewer(palette="RdBu")+
  theme(plot.title = element_text(size=16, color='Darkblue', face='bold', hjust = 0.5))
```



El gráfico de supervivencia en función del tipo de cáncer que la supervivencia media esperada en pacientes con PAAD es inferior a la observada en otros tipos de cáncer, lo cual concuerda con la literatura de pacientes con adenocarcinoma ductal de páncreas (PDAC). Singh et. al (https://doi.org/10.5114/pg.2024.141834) señalan que los pacientes con PDAC presentan el peor pronóstico, con una tasa de supervivencia a cinco años que apenas alcanza el 12–13 %. Y que en cambio, otros tipos de cáncer gastrointestinal como el colorrectal, presentan un pronóstico más favorable, en parte gracias a los programas de detección precoz y al desarrollo de terapias más específicas y eficaces.

Para identificar diferencias estadísticamente significativas entre tipos de cáncer y el tiempo de supervivencia de pacientes, se implementan a continuación las siguientes pruebas: test ANOVA, validación de supuestos de normalidad de residuos del test ANOVA, test Kruskal-Wallis y análisis de supervivencia Kaplan-Mier.

Test ANOVA A continuación, se implementará un test ANOVA asumiendo normalidad, posteriormente se evaluará la normalidad de los residuos del modelo con el propósito de analizar si hay diferencias estadísticamente significativas entre tipos de cáncer.

```
# Eliminar NAs en variables de interés y crear un nuevo dataframe
df_anova <- df_gastro[!is.na(df_gastro$os_time) & !is.na(df_gastro$cancer_type), ]

# Implementar test ANOVA variables os_time y cancer_type
modelo_anova <- aov(os_time ~ cancer_type, data = df_anova)
summary(modelo_anova)
```

```
##               Df    Sum Sq Mean Sq F value   Pr(>F)
## cancer_type    4  17419280 4354820   10.36 3.12e-08 ***
## Residuals   1052 442183622  420327
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De los resultados observamos que el valor-p es 3.12e-08 ($p < 0.05$), lo que indica que hay diferencias significativas entre al menos dos grupos. Además, el F-value es 10.36 (> 1), cuanto mayor sea este valor, mayor evidencia tenemos de diferencias significativas entre grupos. Dado que el ANOVA es significativo, a continuación identificaremos los grupos qué difieren entre sí con la prueba post-hoc TukeyHSD.

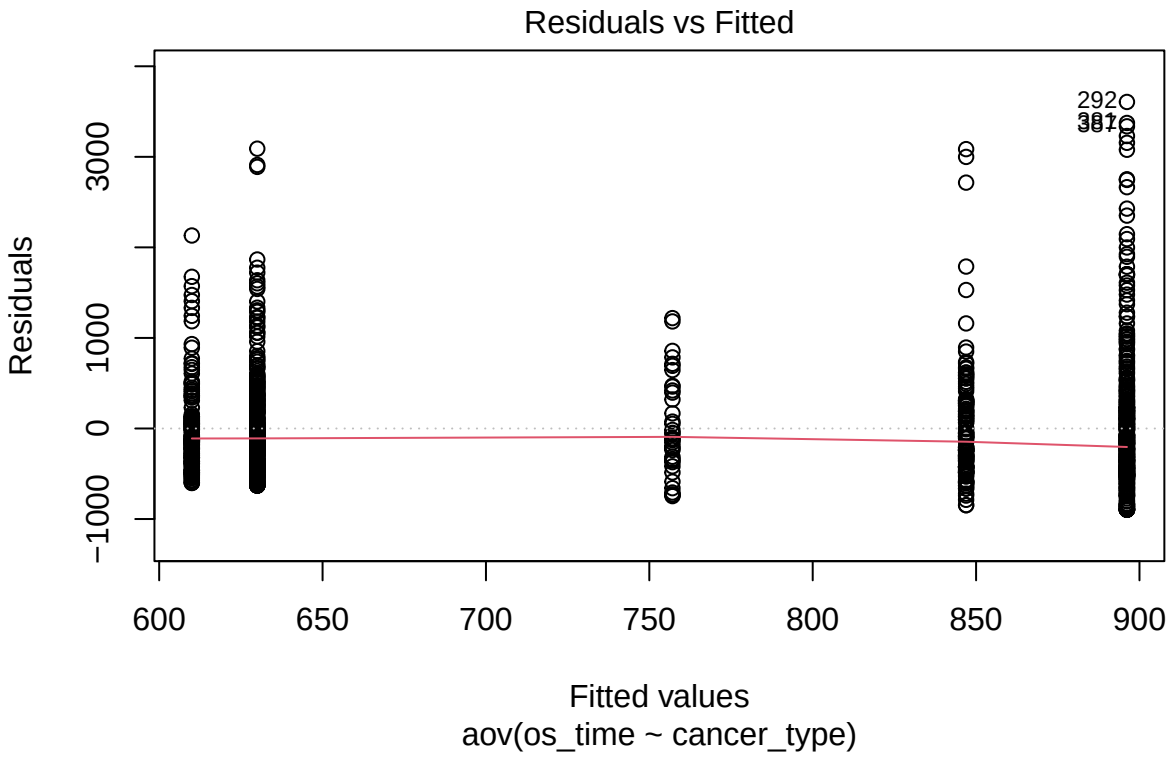
```
# Aplicar test TukeyHSD
TukeyHSD(modelo_anova)
```

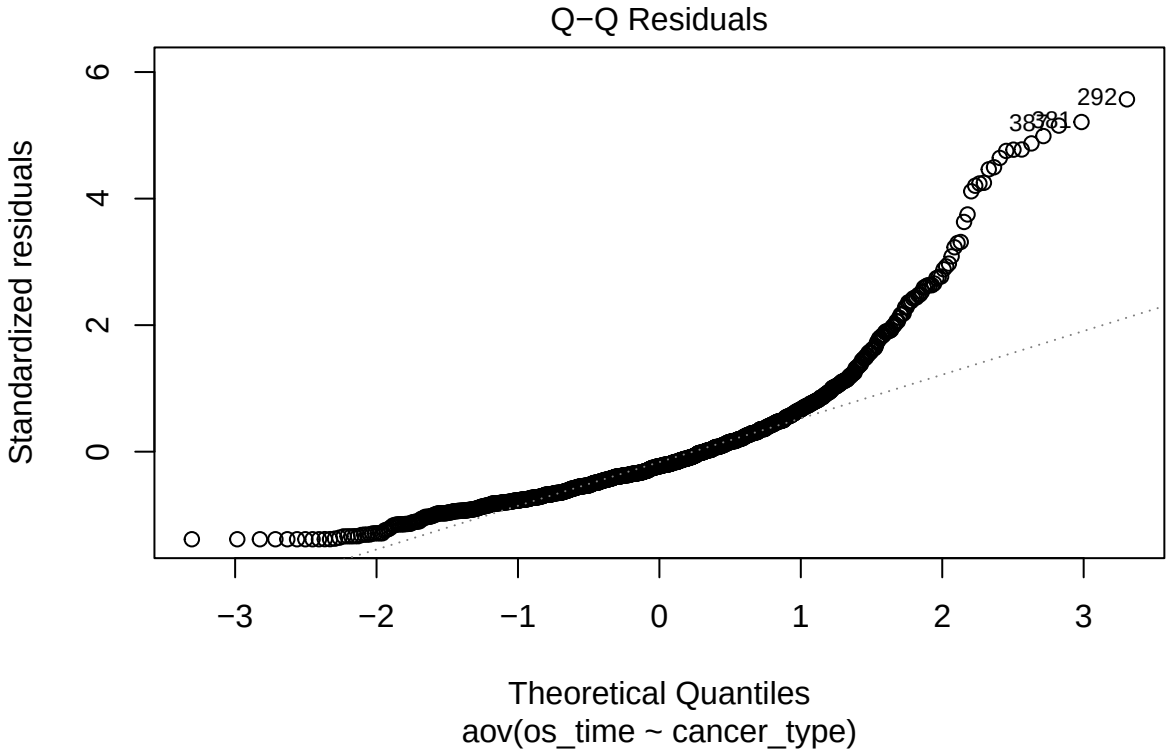
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = os_time ~ cancer_type, data = df_anova)
##
## $cancer_type
##                 diff       lwr        upr     p adj
## COAD-CHOL  139.08982 -161.7348  439.91445 0.7137666
## PAAD-CHOL -147.08580 -469.7097  175.53807 0.7244215
## READ-CHOL   89.86958 -235.1946  414.93381 0.9430886
## STAD-CHOL -127.06130 -430.0807  175.95811 0.7820316
## PAAD-COAD -286.17562 -457.6440 -114.70723 0.0000559
## READ-COAD  -49.22024 -225.2373  126.79680 0.9407864
## STAD-COAD -266.15112 -397.0557 -135.24654 0.0000003
## READ-PAAD  236.95538   25.8329  448.07786 0.0188148
## STAD-PAAD   20.02450 -155.2659  195.31487 0.9979407
## STAD-READ -216.93088 -396.6732  -37.18855 0.0089008
```

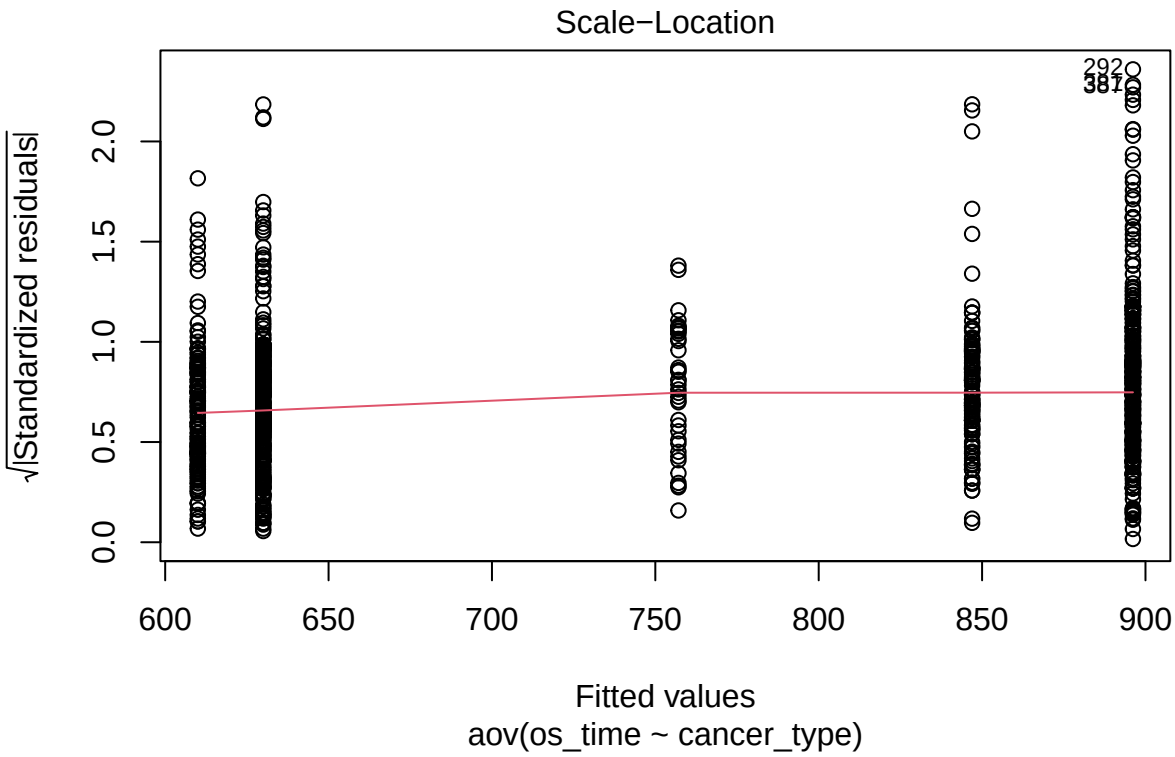
De los resultados obtenidos de la prueba post-hoc `TukeyHSD`, los p-valor ajustados 0.0000003, 0.0000559, 0.0089008, y 0.0188148 (todos ellos < 0.05) muestran que la mayor diferencia significativa está entre los grupos STAD-COAD, PAAD-COAD, STAD-READ y READ-PAAD, respectivamente.

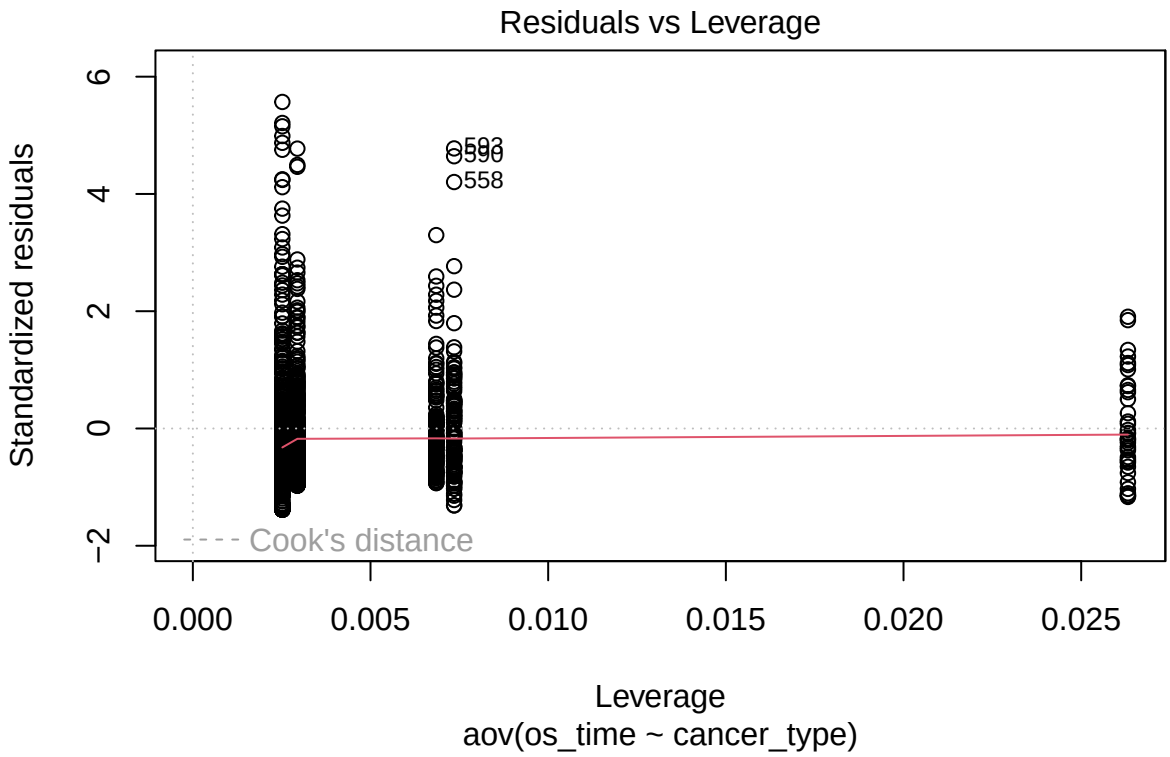
Validación de supuestos de normalidad de residuos del modelo Para validar los resultados del ANOVA, verificaremos los siguientes supuestos: normalidad de residuos, homocedasticidad e independencia.

```
# Validar test ANOVA
plot(modelo_anova)
```


Residuals vs Fitted
Residuals
−1000
0
1000
3000
600
650
700
750
800
850
900
Fitted values
aov(os_time ~ cancer_type)


Q−Q Residuals
Standardized residuals
0
2
4
6
292
−3
−2
−1
0
1
2
3
Theoretical Quantiles
aov(os_time ~ cancer_type)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
600
650
700
750
800
850
900
Fitted values
aov(os_time ~ cancer_type)



Normalidad

Observando la gráfica Q-Q residuals, residuos entre cuantiles (Quantile-Quantile), podemos comparar los cuantiles teóricos de una distribución normal con los cuantiles reales de los residuos del modelo. En nuestro caso, los puntos tienden a alinearse cerca de la línea diagonal, sobretodo en la parte intermedia. No obstante, hay evidencia de asimetría en la cola derecha donde se observa una curva en forma de "S". Este comportamiento puede afectar la validez de los p-valores y las comparaciones post-hoc.

Homocedasticidad

El supuesto de homocedasticidad requiere que la varianza de los residuos sea constante para todos los valores ajustados del modelo para cada grupo. Del gráfico Scale-Location se observa que no hay una relación evidente entre los residuos y los valores ajustados (la media de cada grupo). La banda es casi horizontal y plana. Así que podemos asumir homogeneicidad de varianzas.

Algunas ideas para contestar esta pregunta fueron tomadas del post **ANOVA in R** sección Another method to test normality and homogeneity.

Independencia

Dado que los datos no son temporales, podemos asumir a priori que no hay autocorrelación. La independencia la asumimos por diseño.

Test Kruskal-Wallis Si asumimos que el supuesto de normalidad no se cumple, podemos aplicar las siguientes pruebas no parámetricas `Kruskal-Wallis` y prueba post-hoc `Dunn` para validar nuestros resultados.

```
# Aplicar test Kruskal-Wallis
kruskal.test(os_time ~ cancer_type, data = df_anova)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  os_time by cancer_type
## Kruskal-Wallis chi-squared = 50.438, df = 4, p-value = 2.925e-10
```

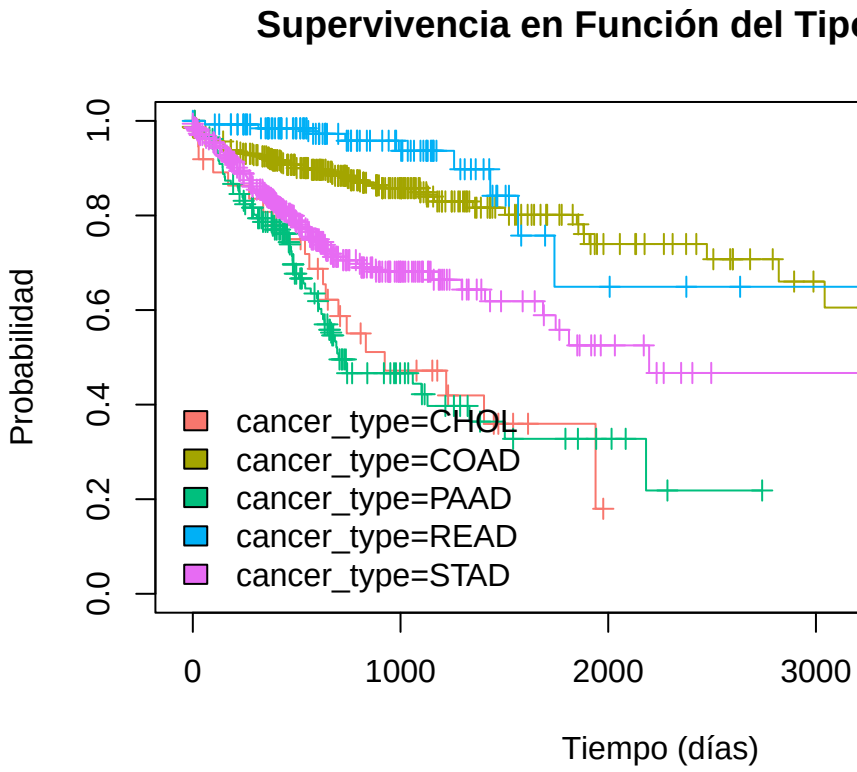
```
# Aplicar post-hoc test Dunn
dunn.test(df_anova$os_time, df_anova$cancer_type, method = "bonferroni")
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 50.4378, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |       CHOL       COAD       PAAD       READ
## ---------+-------------------------------------------
##     COAD |  -0.670817
##          |     1.0000
##          |
##     PAAD |   1.809028   4.580635
##          |     0.3522    0.0000*
##          |
##     READ |  -0.718359  -0.180178  -3.870495
##          |     1.0000     1.0000    0.0005*
##          |
##     STAD |   1.892153   5.921548  -0.058624   4.489049
##          |     0.2924    0.0000*     1.0000    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

Las pruebas no paramétricas efectuadas confirman que los siguientes grupos tienen diferencias estadísticamente significativas: PAAD-COAD, READ-PAAD, STAD-COAD, y STAD-READ.

```
# Implementar Kaplan-Meier por tipo de cancer
# Crear objeto de supervivencia
survival_obj <- Surv(time = df_gastro$os_time, event=df_gastro$os_event)

# Ajustar el model Kaplan-Meier por género
ajuste_cancer_type <- survfit(survival_obj ~ cancer_type, type="kaplan-meier", data=df_gastro)

# Para evitar nombres arbitrarios para el tipo de cáncer necesitamos utilizar los nombres de grupo que
nombres_tipos <- names(ajuste_cancer_type$strata)

# Visualizar gráfico
plot(ajuste_cancer_type, ylab = "Probabilidad", xlab='Tiempo (días)',
```

```
     mark.time = TRUE, col = hue_pal ()(length(nombres_tipos)), main= "Supervivencia en Función del Tipo

legend("bottomleft", legend = nombres_tipos,
       fill = hue_pal ()(length(nombres_tipos)), bty = "n")
```



Estudios de supervivencia (Kaplan-Meier)

```
# Comparamos la supervivencia entre categorías de tipo de cancer
comparar_cancer_type<- survdiff(survival_obj ~ cancer_type, data=df_gastro)
comparar_cancer_type
```

```
## Call:
## survdiff(formula = survival_obj ~ cancer_type, data = df_gastro)
##
## n=1057, 251 observations deleted due to missingness.
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## cancer_type=CHOL  38       20     8.76      14.4     15.01
## cancer_type=COAD 397       57   101.75      19.7     34.75
## cancer_type=PAAD 146       66    28.17      50.8     57.88
## cancer_type=READ 136        9    33.79      18.2     21.23
## cancer_type=STAD 340       87    66.53       6.3      8.79
##
##  Chisq= 111  on 4 degrees of freedom, p= <2e-16
```

Los resultados obtenidos por las pruebas de supervivencia Kaplan-Mier muestran también diferencias significativas en la supervivencia según el tipo de cáncer gastrointestinal, con un valor de Chi-cuadrado de 111,

4 grados de libertad y un valor p inferior a 2e-16.

Tal como se ha señalado previamente en la literatura, los pacientes con adenocarcinoma de páncreas (PAAC) y colangiocarcinoma presentan los peores pronósticos. En nuestro análisis, hemos confirmado que los pacientes con cáncer de páncreas y vías biliares presentan un número de fallecimientos claramente superior al esperado, lo que refleja una supervivencia global significativamente más baja en comparación con otros tipos tumorales como el cáncer de colon o recto.

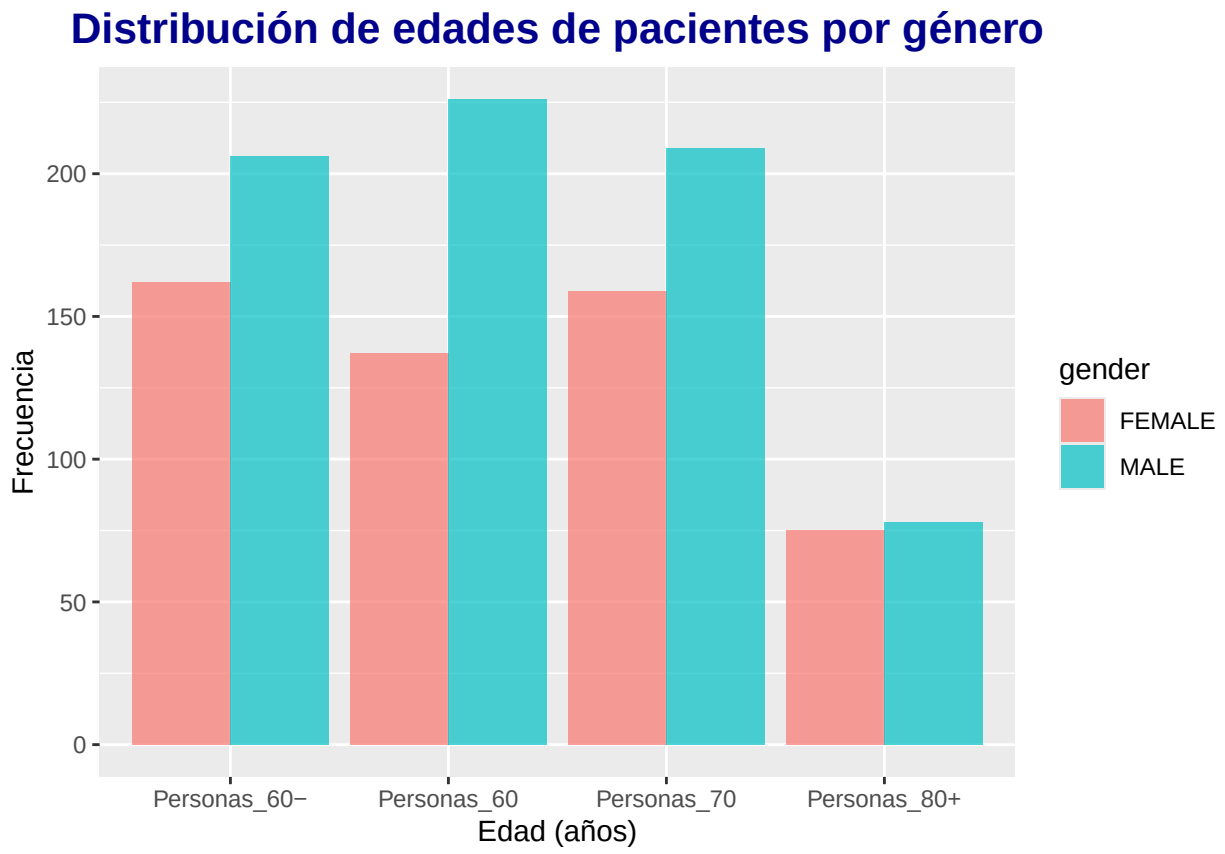
Pregunta 3: ¿Qué diferencias en supervivencia se observan según el género y la edad de los pacientes?

Antes de implementar el análisis de supervivencia por edades, es recomendable realizar la categorización de la variable `age_at_initial_pathologic_diagnosis` procurando una repartición equilibrada entre grupos.

```
# Categorizar la variable age_at_initial_pathologic_diagnosis
df_gastro$age_category <- cut(df_gastro$age_at_initial_pathologic_diagnosis,
                      breaks = c(-Inf,59,69,79,Inf),
                      labels = c("Personas_60-","Personas_60","Personas_70","Personas_80+"),
                      include.lowest = FALSE)

# Verificar distribución de categorias
table(df_gastro$age_category)
```

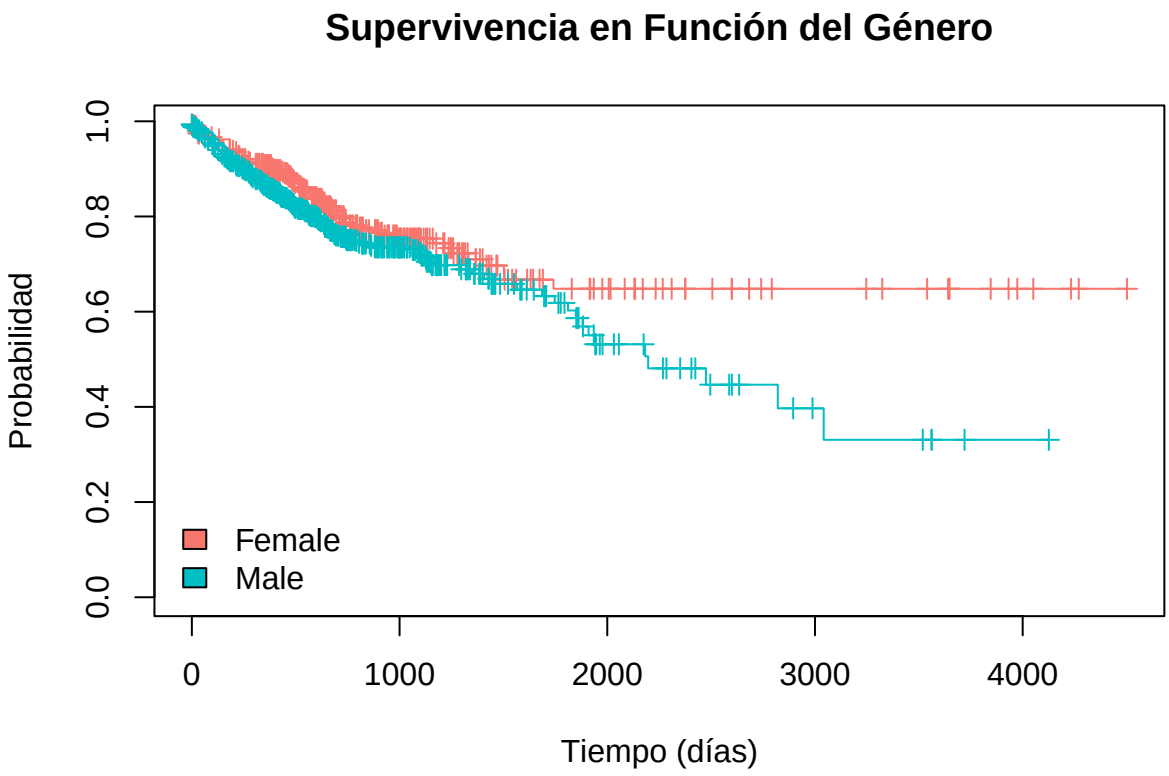
```
##
## Personas_60-  Personas_60  Personas_70 Personas_80+
##          368          363          368          153
```

```
# Distribucion de edades de pacientes por género
ggplot(data=subset(df_gastro, !is.na(age_category)), aes(x= age_category, fill = gender))+
  geom_bar(position = "dodge", alpha = 0.7)+
  labs(title= "Distribución de edades de pacientes por género",
       x= "Edad (años)",
       y= "Frecuencia")+
  theme(plot.title = element_text(size=16, color='Darkblue', face='bold', hjust = 0.5))
```



Estudios de supervivencia (Kaplan-Meier) Los estudios de supervivencia de pacientes en función del género y la edad basados en el modelo Kaplan-Meier se muestran a continuación. Para su implementación se consultaron los siguientes recursos: Survival analysis with low-dimensional input data, Analysis of Cancer Genome Atlas in R y The R Book, 3rd Edition by Elinor Jones, Simon Harden, Michael J. Crawley, cap. 15 Survial Analysis.

```
# Implementar Kaplan-Meier por género
# Crear objeto de supervivencia
survival_obj <- Surv(time = df_gastro$os_time, event=df_gastro$os_event)

# Ajustar el model Kaplan-Meier por género
ajuste_genero <- survfit(survival_obj ~ gender, type="kaplan-meier", data=df_gastro)

# Visualizar gráfico
plot(ajuste_genero, ylab = "Probabilidad", xlab='Tiempo (días)',
     mark.time = TRUE, col = hue_pal ()(2)[1:2], main= "Supervivencia en Función del Género")

legend("bottomleft", legend = c("Female", "Male"),
       fill = hue_pal ()(2)[1:2], bty = "n")
```

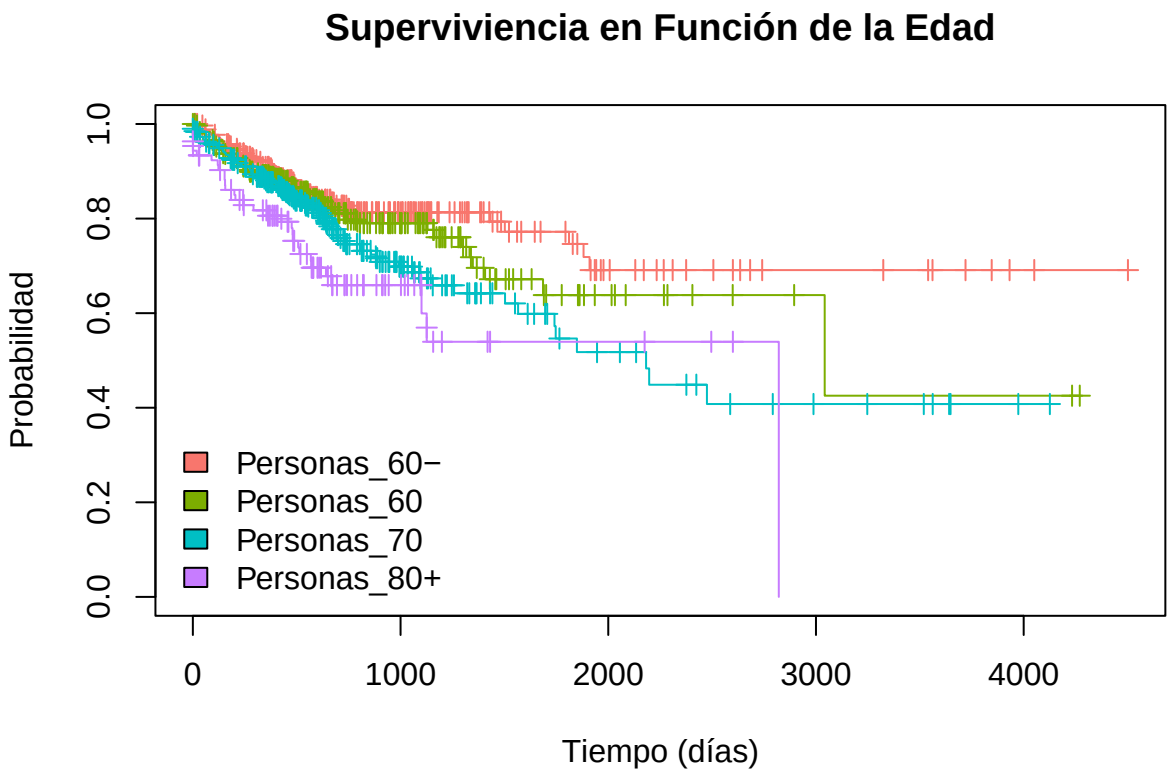



```
# Comparamos la supervivencia entre géneros
comparar_generos <- survdiff(survival_obj ~ gender, data=df_gastro)
comparar_generos
```

```
## Call:
## survdiff(formula = survival_obj ~ gender, data = df_gastro)
##
## n=1057, 251 observations deleted due to missingness.
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## gender=FEMALE 433       86      102      2.54      4.44
## gender=MALE   624      153      137      1.89      4.44
##
##  Chisq= 4.4  on 1 degrees of freedom, p= 0.04
```

En relación a la supervivencia y el género, en el gráfico de Kaplan-Meier, la curva correspondiente a las mujeres alcanza un valor estable de aproximadamente 0.65 a los 1.800 días, lo que indica que alrededor del 65% de ellas sobreviven más allá de ese tiempo. En contraste, la curva de los hombres muestra una caída más prolongada, estabilizándose temporalmente en distintos puntos, y descendiendo hasta aproximadamente 0.35 a los 3.000 días. Esto implica que solo el 35% de los hombres alcanzan ese umbral temporal de supervivencia. El test de log-rank ($Chi^2 = 4.4$, $p = 0.04$) respalda esta inspección visual, donde las diferencias observadas son estadísticamente significativas. Las mujeres tienen menos eventos (muertes) de los esperados, lo que sugiere una mejor evolución clínica en comparación con los hombres.

```r
# Implementar Kaplan-Meier por edad
# Ajustar el model Kaplan-Meier por edad
ajuste_edad <- survfit(survival_obj ~ age_category, type="kaplan-meier", data=df_gastro)

# Visualizar gráfico
plot(ajuste_edad, ylab = "Probabilidad", xlab='Tiempo (días)',
     mark.time = TRUE, col = hue_pal ()(4)[1:4], main= "Superviviencia en Función de la Edad")

legend("bottomleft", legend = c("Personas_60-","Personas_60","Personas_70","Personas_80+"),
       fill = hue_pal ()(4)[1:4], bty = "n")
```




```r
# Comparamos la supervivencia entre categorías de edades
comparar_edades <- survdiff(survival_obj ~ age_category, data=df_gastro)
comparar_edades
```

```
## Call:
## survdiff(formula = survival_obj ~ age_category, data = df_gastro)
##
## n=1016, 292 observations deleted due to missingness.
##
##                             N Observed Expected (O-E)^2/E (O-E)^2/V
## age_category=Personas_60- 311       51     70.2     5.268     7.789
## age_category=Personas_60  298       58     64.7     0.695     0.989
## age_category=Personas_70  298       78     65.3     2.482     3.540
## age_category=Personas_80+ 109       32     18.8     9.290    10.197
```

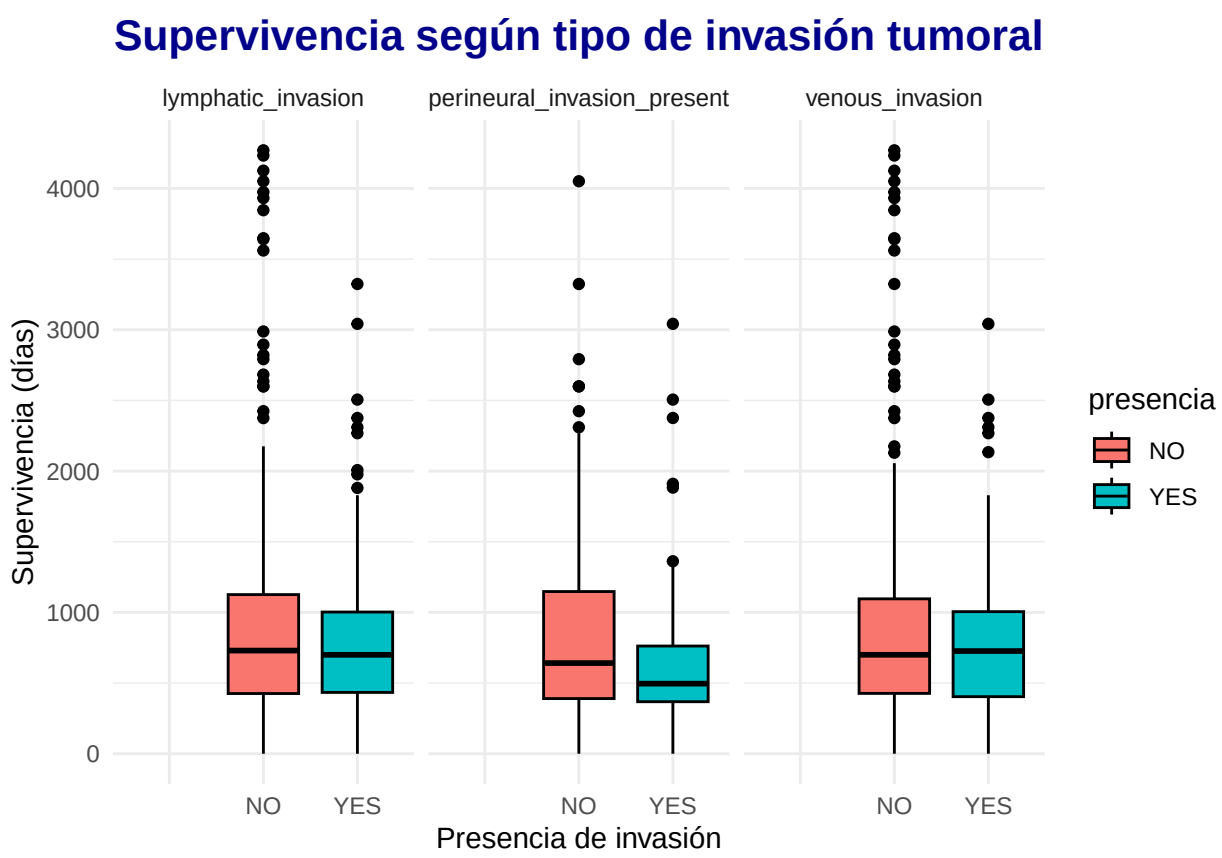
```
##
##  Chisq= 17.8  on 3 degrees of freedom, p= 5e-04
```

En relación a la supervivencia y la edad, en el gráfico de Kaplan-Meier, los grupos más jóvenes mantienen curvas de supervivencia más altas y sostenidas a lo largo del tiempo, mientras que las curvas de los mayores de 70 y 80 años muestran un descenso más pronunciado. El test de log-rank ($Chi^2 = 17.8$, $gl = 3$, $p = 0.0005$) confirma que entre los grupos etarios existen diferencias estadísticamente significativas. En particular, se observa que el grupo de mayores de 80 años tiene una mortalidad observada notablemente más alta que la esperada (32 observados vs 18.8 esperados). Por el contrario, el grupo de personas menores de 60 años presenta una mortalidad observada más baja que la esperada (51 vs 70.2), indicando una mejor supervivencia relativa.

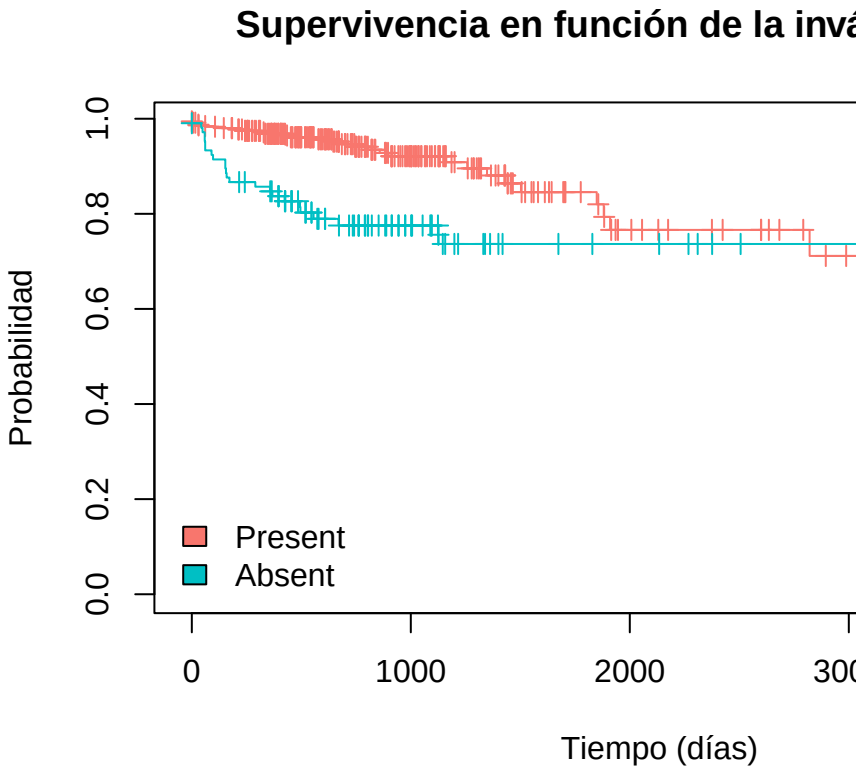
Pregunta 4: ¿Cuál es el impacto de la invasión venosa, linfática o perineural sobre la supervivencia?

```r
# Reorganizar los datos
df_long <- df_gastro %>%
  select(os_time, perineural_invasion_present, venous_invasion, lymphatic_invasion) %>%
  pivot_longer(
    cols = c(perineural_invasion_present, venous_invasion, lymphatic_invasion),
    names_to = "tipo_invasion",
    values_to = "presencia"
  ) %>%
  filter(!is.na(presencia))

# Crear el gráfico combinado
ggplot(df_long, aes(x = presencia, y = os_time, fill = presencia)) +
  geom_boxplot(color = "black") +
  facet_wrap(~ tipo_invasion, scales = "free_x") +
  labs(
    title = "Supervivencia según tipo de invasión tumoral",
    x = "Presencia de invasión",
    y = "Supervivencia (días)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, color = 'Darkblue', face = 'bold', hjust = 0.5)
  )
```

```
## Warning: Removed 195 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```




```r
# Implementar Kaplan-Meier invásion venosa
# Crear objeto de supervivencia
survival_obj <- Surv(time = df_gastro$os_time, event=df_gastro$os_event)

# Ajustar el model Kaplan-Meier por género
ajuste_invasion_venosa<- survfit(survival_obj ~ venous_invasion, type="kaplan-meier", data=df_gastro)

# Visualizar gráfico
plot(ajuste_invasion_venosa, ylab = "Probabilidad", xlab='Tiempo (días)',
     mark.time = TRUE, col = hue_pal ()(2)[1:2], main= "Supervivencia en función de la invásion venosa")

legend("bottomleft", legend = c("Present", "Absent"),
       fill = hue_pal ()(2)[1:2], bty = "n")
```



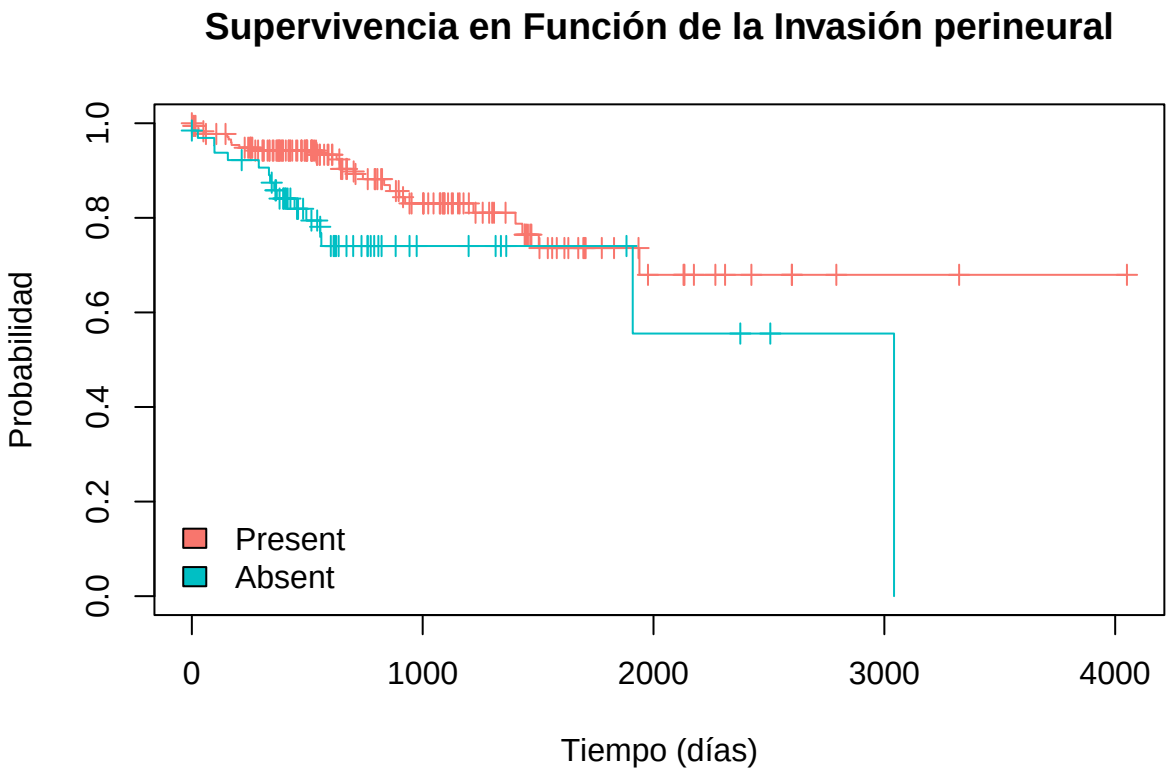
Estudios de supervivencia (Kaplan-Meier)

```
# Comparamos la supervivencia entre pacientes con o sin invásion
comparar_invasion_venosa <- survdiff(survival_obj ~ venous_invasion, data=df_gastro)
comparar_invasion_venosa
```

```
## Call:
## survdiff(formula = survival_obj ~ venous_invasion, data = df_gastro)
##
## n=467, 841 observations deleted due to missingness.
##
##                         N Observed Expected (O-E)^2/E (O-E)^2/V
## venous_invasion=NO  361       29     42.1      4.08        20
## venous_invasion=YES 106       24     10.9     15.77        20
##
##  Chisq= 20  on 1 degrees of freedom, p= 8e-06
```

```
# Implementar Kaplan-Meier invasion perineural
# Crear objeto de supervivencia
survival_obj <- Surv(time = df_gastro$os_time, event=df_gastro$os_event)

# Ajustar el model Kaplan-Meier por género
ajuste_invasion_perineural<- survfit(survival_obj ~ perineural_invasion_present, type="kaplan-meier", da

# Visualizar gráfico
plot(ajuste_invasion_perineural, ylab = "Probabilidad", xlab='Tiempo (días)',
     mark.time = TRUE, col = hue_pal ()(2)[1:2], main= "Supervivencia en Función de la Invasión perineu
```

```
legend("bottomleft", legend = c("Present", "Absent"),
       fill = hue_pal ()(2)[1:2], bty = "n")
```

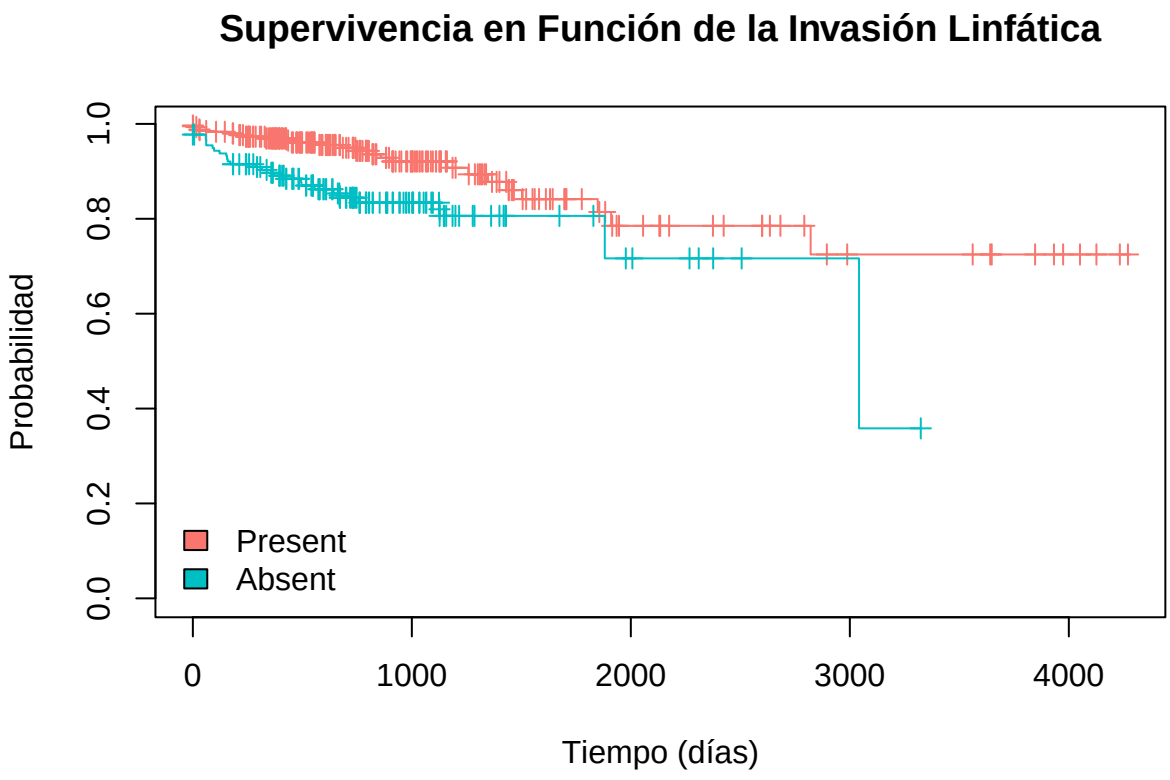



```
# Comparamos la supervivencia entre los pacientes con y sin invasión perineural
comparar_invasion_perineural <- survdiff(survival_obj ~ perineural_invasion_present, data=df_gastro)
comparar_invasion_perineural
```

```
## Call:
## survdiff(formula = survival_obj ~ perineural_invasion_present,
##     data = df_gastro)
##
## n=247, 1061 observations deleted due to missingness.
##
##                                 N Observed Expected (O-E)^2/E (O-E)^2/V
## perineural_invasion_present=NO  182       25    31.74      1.43      6.43
## perineural_invasion_present=YES  65       16     9.26      4.91      6.43
##
##  Chisq= 6.4  on 1 degrees of freedom, p= 0.01
```

```
# Implementar Kaplan-Meier Invasión linfática
# Crear objeto de supervivencia
survival_obj <- Surv(time = df_gastro$os_time, event=df_gastro$os_event)

# Ajustar el model Kaplan-Meier por género
```

```r
ajuste_invasion_linfatica<- survfit(survival_obj ~ lymphatic_invasion, type="kaplan-meier", data=df_gas

# Visualizar gráfico
plot(ajuste_invasion_linfatica, ylab = "Probabilidad", xlab='Tiempo (días)',
     mark.time = TRUE, col = hue_pal ()(2)[1:2], main= "Supervivencia en Función de la Invasión Linfáti

legend("bottomleft", legend = c("Present", "Absent"),
       fill = hue_pal ()(2)[1:2], bty = "n")
```




```r
# Comparamos la supervivencia entre los pacientes con y sin invasión perineural
comparar_invasion_linfatica <- survdiff(survival_obj ~ lymphatic_invasion, data=df_gastro)
comparar_invasion_linfatica
```

```
## Call:
## survdiff(formula = survival_obj ~ lymphatic_invasion, data = df_gastro)
##
## n=483, 825 observations deleted due to missingness.
##
##                           N Observed Expected (O-E)^2/E (O-E)^2/V
## lymphatic_invasion=NO   305       25       36      3.37      10.3
## lymphatic_invasion=YES 178       29       18      6.75      10.3
##
##  Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

El análisis de supervivencia según variables histopatológicas reveló diferencias significativas asociadas a la

presencia de invasión venosa, linfática y perineural. Estas tres características son conocidas por reflejar una mayor agresividad tumoral, y nuestros resultados lo confirmaron estadísticamente.

En primer lugar, la invasión venosa mostró una asociación altamente significativa con la supervivencia (Chi-cuadrado = 20, p = 8e-06), con un número de fallecimientos muy superior al esperado en pacientes con invasión positiva. De forma similar, la invasión linfática también se relacionó con peor pronóstico (Chi-cuadrado = 10,3, p = 0,001), al observarse una mayor mortalidad en el grupo con invasión.

Por último, la invasión perineural se asoció de forma significativa con una reducción en la supervivencia (Chi-cuadrado = 6,4, p = 0,01), aunque con una cohorte más limitada en tamaño.

Estos resultados respaldan el valor clínico de estas variables como posibles marcadores pronósticos negativos en pacientes con cáncer gastrointestinal. Además, resaltan la importancia de incorporarlas en la estratificación del riesgo y en la toma de decisiones terapéuticas.

3.2 Función personalizada y ejercicios de probabilidad

3.2.1 Función personalizada

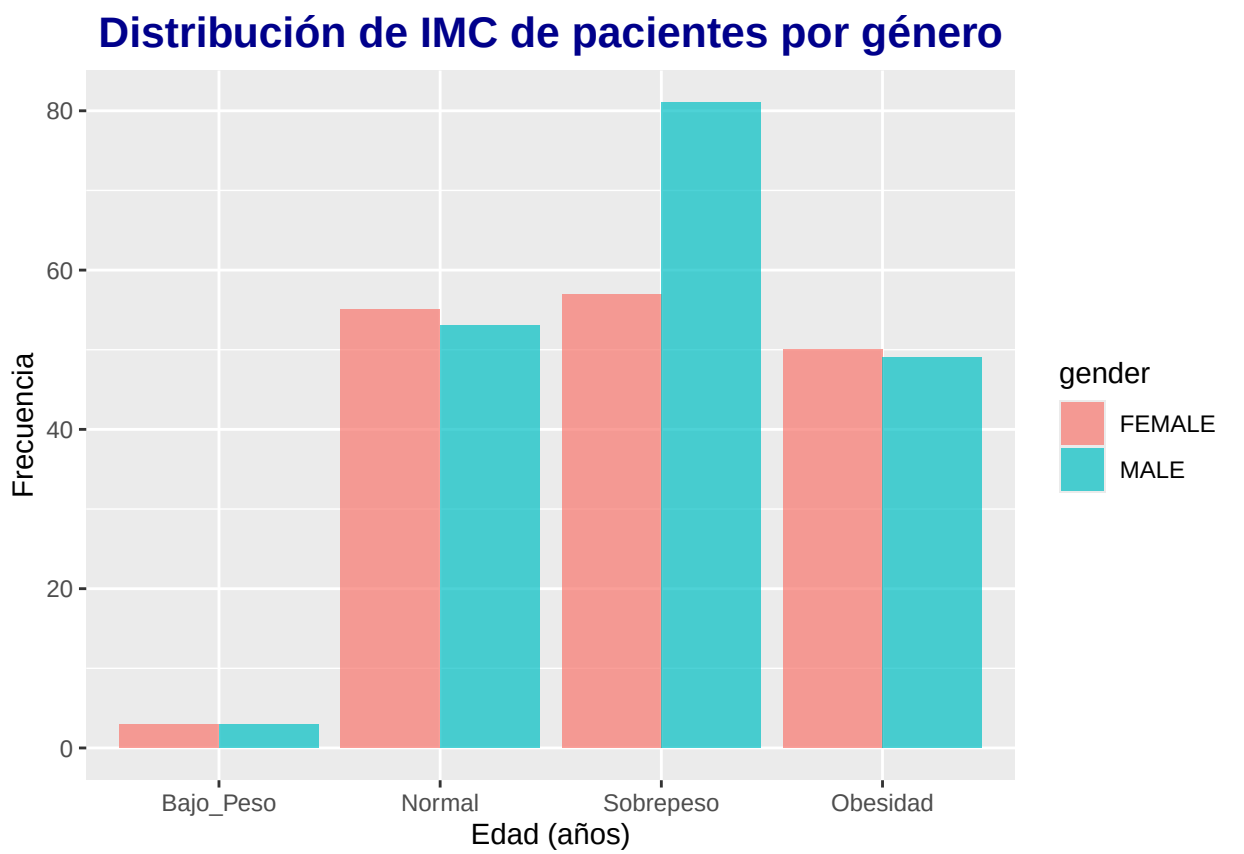
El IMC (Índice de Masa Corporal) es un indicador ampliamente utilizado en oncología para evaluar riesgo metabólico, pronóstico y respuesta a tratamientos, conforme al punto de vista del grupo IARC en su reporte Body Fatness and Cancer — Viewpoint of the IARC Working Group. El IMC al momento del diagnóstico puede reflejar comorbilidades (diabetes, hipertensión) que pueden influir en la progresión del cáncer. Es por esto que hemos incluido el IMC como una variable clave para la estratificación de pacientes tomando como referencia las categorías propuestas por el NIH. Para más detalles, consultar Obsedidad y Cáncer.

```
# Función para calcular el IMC de un paciente
calcular_imc <- function(peso, altura){
  imc <- peso / (altura/100)^2
  return (imc)
}

# calcular_imc <- function(peso, altura){
#   tryCatch(
#     {
#       # Convertir en valores numéricos por default
#       #peso      <- as.numeric(peso)
#       #altura    <- as.numeric(altura)
#
#       # Validar rango de valores de entrada.
#       if(altura <= 0.0){
#         stop("Error: La variable altura no debe ser negativa o cero")
#       }
#
#       imc <- peso / (altura^2)
#       return(imc)
#     },
#     error = function(e){
#         message("Error en el cálculo: ", e$message)
#         return(NA)
#     }
#   )
# }

# Crear la variable imc_at_initial_pathologic_diagnosis
```

```r
df_gastro$imc <- calcular_imc(df_gastro$weight, df_gastro$height)

# Categorizar la variable imc según las categorías propuestas por NIH
df_gastro$imc_category <- cut(df_gastro$imc,
                        breaks = c(-Inf,18.5,25.0,30.0,Inf),
                        labels = c("Bajo_Peso","Normal","Sobrepeso","Obesidad"),
                        include.lowest = FALSE)

# Verificar distribución de categorias
table(df_gastro$imc_category)
```

```
##
## Bajo_Peso    Normal Sobrepeso  Obesidad
##         6       108       138        99
```

```r
# Distribucion de IMC de pacientes por género
ggplot(data=subset(df_gastro, !is.na(imc_category)), aes(x= imc_category, fill = gender))+
  geom_bar(position = "dodge", alpha = 0.7)+
  labs(title= "Distribución de IMC de pacientes por género",
       x= "Edad (años)",
       y= "Frecuencia")+
  theme(plot.title = element_text(size=16, color='Darkblue', face='bold', hjust = 0.5))
```



3.2.2 Ejercicios de probabilidad

Esta sección propone tres problemas de probabilidad relacionados al dataset TCGA.

Ejercicio A

En el dataset TCGA personalizado (df_gastro), aproximadamente el 14.15% de los pacientes tienen cáncer de páncreas, esto conforme a la variable `cancer_type`. Si tomamos una muestra aleatoria de 100 pacientes:

- ¿Cuál es la probabilidad de que exactamente 5 pacientes tengan cáncer de páncreas?
- ¿Cuál es la probabilidad de que a lo más 10 pacientes tengan cáncer de páncreas?

```
# Verificar distribución de categorias por tipo de cáncer
frecuencias <- table(df_gastro$cancer_type)
frecuencias
```

```
##
##      CHOL COAD PAAD READ STAD
##    1   48  459  185  171  443
```

```
# Verificar porcentaje de distribución de categorías por tipo de cáncer
# porcentajes <- prop.table(frecuencias) * 100
porcentajes <- round(prop.table(frecuencias) * 100, 2)
porcentajes
```

```
##
##        CHOL  COAD  PAAD  READ  STAD
##  0.08  3.67 35.12 14.15 13.08 33.89
```

```
# Describir datos para calcular P(X = 5)
n <- 100
p <- 0.1415
x <- 5

# Aplicar distribución binomial para P(X = 5)
paciente_5 <- dbinom(x,n,p)

print(paste("La probabilidad de que exactamente 5 pacientes tengan cáncer de páncreas es:",
            round(paciente_5,4)*100,"%"))
```

```
## [1] "La probabilidad de que exactamente 5 pacientes tengan cáncer de páncreas es: 0.22 %"
```

```
# Describir datos para calcular P(X <= 10)
n <- 100
p <- 0.1415
x <- 10

# Aplicar distribución binomial para P(X <= 10)
paciente_10 <- pbinom(x, n, p)

print(paste("La probabilidad de que un máximo de 10 pacientes tengan cáncer de páncreas es:",
            round(paciente_10, 4)*100,"%"))
```

```
## [1] "La probabilidad de que un máximo de 10 pacientes tengan cáncer de páncreas es: 14.62 %"
```

Ejercicio B

De los pacientes con invasión perineural ¿cuál es la probabilidad de que estos pacientes tengan un estadio patológico avanzado, igual o superior a tipo III?

Este es un problema de probabilidad condicional donde se nos pide: P(Estadio >= III | Invasión perineural = YES)

```
# Visualizar variables categóricas de interés
print("Distribución de categorías para la variable perineural_invasion_present")
```

```
## [1] "Distribución de categorías para la variable perineural_invasion_present"
```

```
table(df_gastro$perineural_invasion_present)
```

```
##
##      NO YES
##   1 206  71
```

```
print("Distribución de categorías para la variable stage_event_pathologic_stage")
```

```
## [1] "Distribución de categorías para la variable stage_event_pathologic_stage"
```

```
table(df_gastro$stage_event_pathologic_stage)
```

```
##
##         I   IA   IB   II  IIA  IIB  IIC  III IIIA IIIB IIIC   IV  IVA  IVB
##    1  131   22   56   82  248  193    2   36   94  149   93  115   27    7
```

```
# Eliminar NAs en variables de interés y crear un nuevo dataframe
df_estate_peri <- df_gastro[!is.na(df_gastro$perineural_invasion_present) &
                      !is.na(df_gastro$stage_event_pathologic_stage), ]
```

Para estudiar la relación entre los grupos de variables categóricas, el siguiente paso es construir una tabla de contingencia de dimensión 14 x 2, donde se mostrarán los conteos para cada combinación invasion_perineural–estadio, resultando en 28 frecuencias observadas.

```
# Crear tabla de contingencia
tabla_conti <- table(df_estate_peri$perineural_invasion_present,
               df_estate_peri$stage_event_pathologic_stage)
print("Tabla de contingencia de variables categóricas")
```

```
## [1] "Tabla de contingencia de variables categóricas"
```

```
tabla_conti
```

```
##
##           I IA IB II IIA IIB IIC III IIIA IIIB IIIC IV IVA IVB
##        1  0  0  0  0   0   0   0   0    0    0    0  0   0   0
##   NO   0 53  1  0  9  52   4   1   7    9   31   14  6  10   4
##   YES  0  8  0  0  4  13   5   0   1    0   13    4 10   6   3
```

```r
# Calcular proporciones para cada grupo de pacientes con invasión perineural
prop_conti <- prop.table(tabla_conti, margin = 1)  # margin = 1 obtiene proporciones por fila
print("Tabla de proporciones")
```

```
## [1] "Tabla de proporciones"
```

```r
prop_conti
```

```
##
##                           I          IA          IB          II         IIA
##       1.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000
##   NO  0.000000000 0.263681592 0.004975124 0.000000000 0.044776119 0.258706468
##   YES 0.000000000 0.119402985 0.000000000 0.000000000 0.059701493 0.194029851
##
##               IIB         IIC         III        IIIA        IIIB        IIIC
##       0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000
##   NO  0.019900498 0.004975124 0.034825871 0.044776119 0.154228856 0.069651741
##   YES 0.074626866 0.000000000 0.014925373 0.000000000 0.194029851 0.059701493
##
##                IV         IVA         IVB
##       0.000000000 0.000000000 0.000000000
##   NO  0.029850746 0.049751244 0.019900498
##   YES 0.149253731 0.089552239 0.044776119
```

```r
# Obtener la proporción solo para el grupo con invasión perineural = 'YES' y estadios > III
print("Tabla de proporciones para pacientes con invasión perineural = 'YES' y estadios > III")
```

```
## [1] "Tabla de proporciones para pacientes con invasión perineural = 'YES' y estadios > III"
```

```r
prop_conti["YES", c("III", "IIIA", "IIIB", "IIIC", "IV", "IVA", "IVB")]
```

```
##        III       IIIA       IIIB       IIIC         IV        IVA        IVB
## 0.01492537 0.00000000 0.19402985 0.05970149 0.14925373 0.08955224 0.04477612
```

```r
# Obtener la probabilidad total conjunta
prob_conjunta <- sum(prop_conti["YES", c("III", "IIIA", "IIIB", "IIIC", "IV", "IVA", "IVB")])
paste("Probabilidad estadio igual o superior a tipo III:", round(prob_conjunta, 4)*100,"%")
```

```
## [1] "Probabilidad estadio igual o superior a tipo III: 55.22 %"
```

Ejercicio C

La edad media al diagnóstico es de 65.85 años con una desviación estándar de 11.88 años. Suponiendo una distribución normal, ¿cuál es la probabilidad de que un paciente tenga entre 60 y 70 años?

En este problema utilizaremos la distribución normal para obtener $P(60 < X < 70)$

```r
# Describir datos para obtener P(60 < X < 70)
media  <- mean(df_gastro$age_at_initial_pathologic_diagnosis, na.rm = TRUE)
sigma  <- sd(df_gastro$age_at_initial_pathologic_diagnosis, na.rm = TRUE)
```

```r
x1     <- 60
x2     <- 70

# Aplicar distribución normal para obtener la probabilidad P(60 < X < 70)
menor_a_70   <- pnorm(x2, media, sigma)
menor_a_60   <- pnorm(x1, media, sigma)
entre_70_60 <- menor_a_70 - menor_a_60

print(paste("La probabilidad de que un paciente tenga entre 60 y 70 años es:",
            round(entre_70_60, 4)*100,"%"))
```

```
## [1] "La probabilidad de que un paciente tenga entre 60 y 70 años es: 32.55 %"
```

Sección 4: Modelos de aprendizaje automático

Dado que nuestro objetivo es identificar factores asociados a la supervivencia de pacientes oncológicos, optamos por aplicar modelos de aprendizaje supervisado, ya que contamos con una variable dependiente bien definida: la supervivencia global (`os_event`).

Entrenamos un modelo de aprendizaje supervisado utilizando Support Vector Machine (SVM). Para ello, seleccionamos un subconjunto de variables clínicas incluyendo edad al diagnóstico, género, tipo de tumor residual, estadio patológico, número de ganglios positivos, tipo de cáncer y presencia de invasión venosa, linfática o perineural.

Modelo supervisado (SVG)

```r
# Filtramos filas sin valores NA en variables clave
model_df_supervised <- na.omit(df_gastro[, c("age_at_initial_pathologic_diagnosis",
                                "gender", "residual_tumor", "stage_event_pathologic_stage",
                               "number_of_lymphnodes_positive_by_he",
                               "cancer_type",
                               "venous_invasion", "lymphatic_invasion",
                               "perineural_invasion_present",
                               "os_event")])

# Convertimos os_event en factor (para clasificación)
model_df_supervised$os_event <- as.factor(model_df_supervised$os_event)

# Dividimos el conjunto de datos en entrenamiento y prueba
set.seed(123)  # Para reproducibilidad
indices_entrenamiento <- sample(1:nrow(model_df_supervised), 0.5 * nrow(model_df_supervised))
conjunto_entrenamiento <- model_df_supervised[indices_entrenamiento, ]
conjunto_prueba <- model_df_supervised[-indices_entrenamiento, ]

# Cargamos el paquete "kernlab" para SVM
library(kernlab)
```

```
## Warning: package 'kernlab' was built under R version 4.4.1
```

```
##
## Attaching package: 'kernlab'

## The following object is masked from 'package:ggplot2':
##
##     alpha

## The following object is masked from 'package:scales':
##
##     alpha
```

```
# Entrenamos un modelo SVM
modelo_svm <- ksvm(os_event ~ ., data = conjunto_entrenamiento, kernel = "rbfdot", type = "C-svc") #ind

# Realizar predicciones en el conjunto de prueba
predicciones <- predict(modelo_svm, newdata = conjunto_prueba)

# Calcular la matriz de confusión
confusion_matriz <- table(Real = conjunto_prueba$os_event, Predicción =
predicciones)

# Calcular la precisión
precision <- sum(diag(confusion_matriz)) / sum(confusion_matriz)
cat("Precisión del modelo SVM:", precision, "\n")
```

```
## Precisión del modelo SVM: 0.9354839
```

```
# Mostrar la matriz de confusión
confusion_matriz
```

```
##     Predicción
## Real  0  1
##    0 87  0
##    1  6  0
```

Preprocesamiento del conjunto de datos

Se filtraron las filas que presentaban valores NA en las variables seleccionadas. Además, se convirtió la variable `os_event` en un factor para permitir la clasificación. El conjunto resultante fue dividido aleatoriamente en dos subconjuntos: entrenamiento (50%) y prueba (50%).

Entrenamiento del modelo

Se utilizó la función `ksvm()` del paquete `kernlab` para entrenar un modelo SVM con kernel radial `rbfdot`. El modelo se ajustó utilizando los datos de entrenamiento y luego se evaluó su rendimiento sobre el conjunto de prueba.

Resultados

El modelo alcanzó una precisión del 93.5% en el conjunto de prueba. No obstante, la matriz de confusión revela que el modelo no logró clasificar correctamente a ningún paciente fallecido: todos los pacientes que murieron fueron clasificados como vivos. Esto indica que el modelo tiene un sesgo hacia la clase mayoritaria (pacientes vivos), probablemente debido a un fuerte desbalance en la distribución de los pacientes vivos/fallecidos. En otras palabras, aunque el modelo parece preciso en términos generales, su capacidad para detectar pacientes con peor pronóstico es nula. Para corregir este desbalance, haría falta añadir mas pacientes con peor pronóstico.

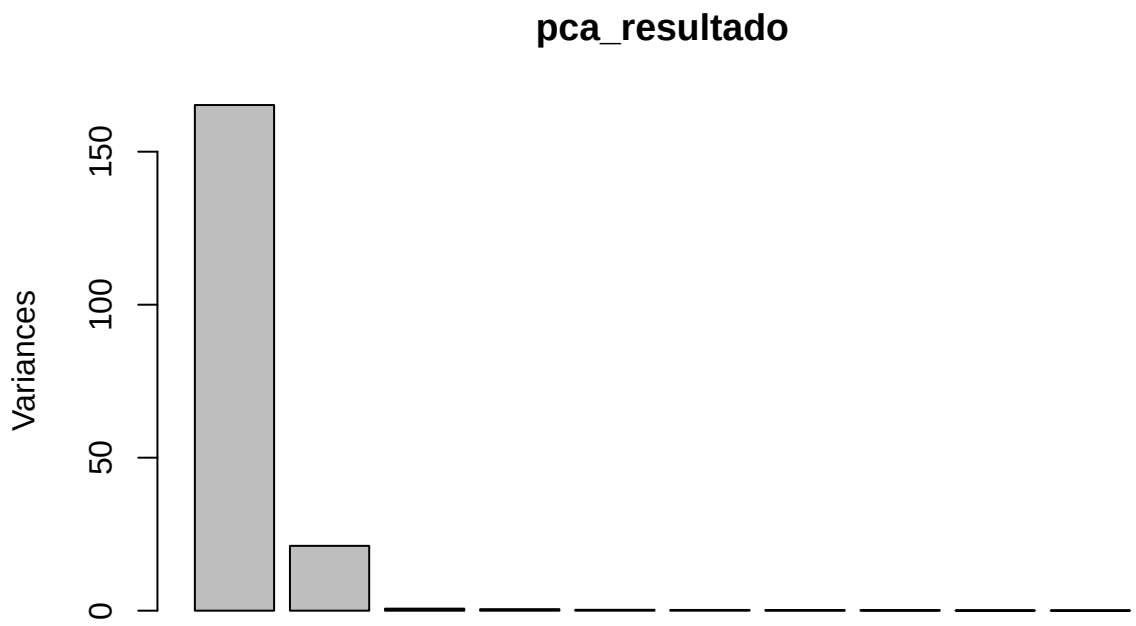
Modelo no supervisado PCA

En esta parte del trabajo usamos un modelo de aprendizaje no supervisado para explorar si los datos clínicos de los pacientes muestran algún patrón o agrupación natural, sin necesidad de decirle al modelo qué tiene que predecir. Para eso usamos un Análisis de Componentes Principales (PCA).

```
# # Filtramos filas sin valores NA en variables clave
model_df_unsupervised <- na.omit(df_gastro[, c("age_at_initial_pathologic_diagnosis",
                                 "gender", "residual_tumor", "stage_event_pathologic_stage",
                                "number_of_lymphnodes_positive_by_he",
                                "cancer_type",
                                "venous_invasion", "lymphatic_invasion",
                                "perineural_invasion_present")])

# Transformamos las variables categoricas en numericas
# Ref. "The model.matrix function" The R Book page 255
df_numeric <- model.matrix(~ gender + age_at_initial_pathologic_diagnosis +
                             residual_tumor +
                             number_of_lymphnodes_positive_by_he +
                             venous_invasion +
                             lymphatic_invasion +
                             perineural_invasion_present +
                             stage_event_pathologic_stage - 1, data = model_df_unsupervised)

# Aplicamos PCA
pca_resultado <- prcomp(df_numeric, center = TRUE)

# Resumen de la varianza explicada por cada variable
summary(pca_resultado)
```

```
## Importance of components:
##                            PC1    PC2     PC3    PC4     PC5     PC6     PC7
## Standard deviation     12.855 4.6032 0.81452 0.7006 0.52639 0.48330 0.43598
## Proportion of Variance  0.875 0.1122 0.00351 0.0026 0.00147 0.00124 0.00101
## Cumulative Proportion   0.875 0.9872 0.99070 0.9933 0.99477 0.99600 0.99701
##                            PC8     PC9    PC10   PC11    PC12   PC13    PC14
## Standard deviation     0.41054 0.29120 0.28441 0.2380 0.22723 0.1941 0.15327
## Proportion of Variance 0.00089 0.00045 0.00043 0.0003 0.00027 0.0002 0.00012
## Cumulative Proportion  0.99790 0.99835 0.99878 0.9991 0.99935 0.9996 0.99968
##                           PC15    PC16    PC17    PC18    PC19    PC20
## Standard deviation     0.14545 0.11274 0.10065 0.07723 0.07598 0.07337
## Proportion of Variance 0.00011 0.00007 0.00005 0.00003 0.00003 0.00003
## Cumulative Proportion  0.99979 0.99986 0.99991 0.99994 0.99997 1.00000
##                             PC21      PC22      PC23      PC24      PC25
## Standard deviation     9.744e-16 8.971e-16 8.971e-16 8.971e-16 8.971e-16
## Proportion of Variance 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  1.000e+00 1.000e+00 1.000e+00 1.000e+00 1.000e+00
##                             PC26      PC27      PC28      PC29
## Standard deviation     8.971e-16 8.971e-16 8.971e-16 8.112e-16
## Proportion of Variance 0.000e+00 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  1.000e+00 1.000e+00 1.000e+00 1.000e+00
```

```r
# Representamos la varianza explicada por cada componente
plot(pca_resultado)
```




```r
cargos <- pca_resultado$rotation
#print(cargos)
```

Resultados PCA

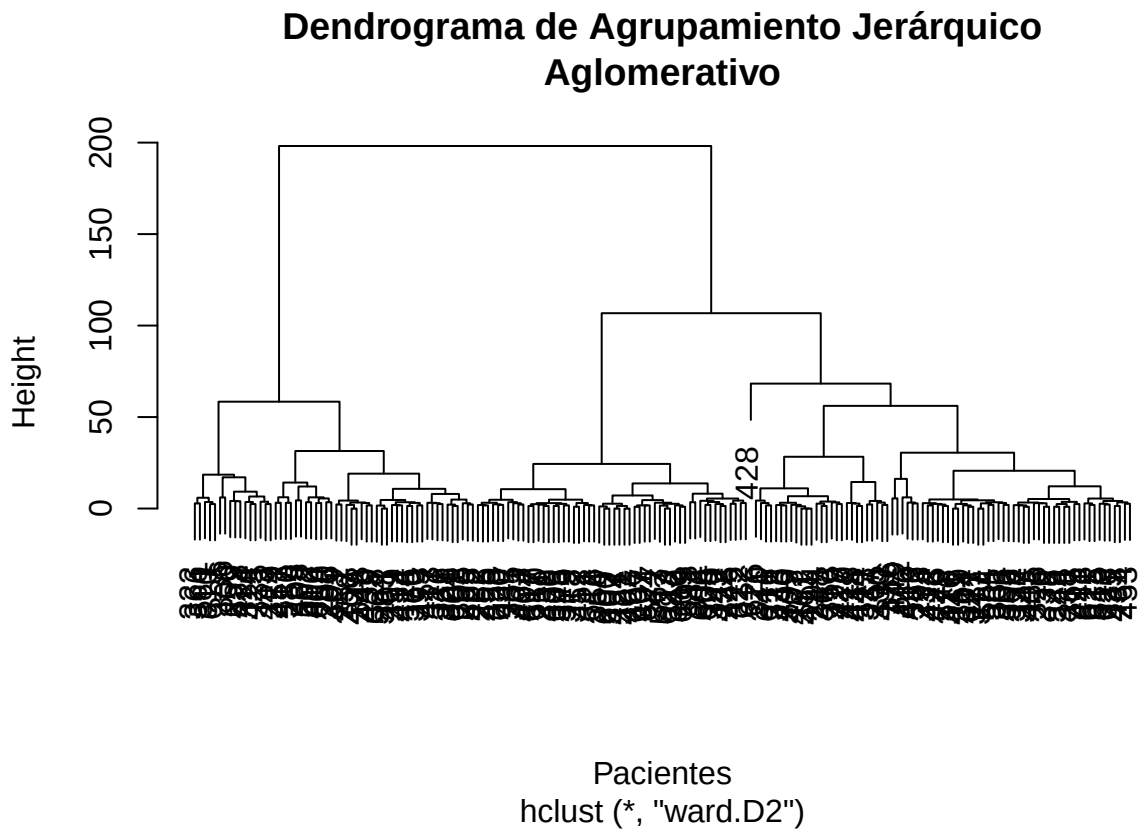
El primer componente (PC1) explica el 87.5% de la varianza total del conjunto de datos, mientras que el segundo componente (PC2) añade un 11.2% adicional. En conjunto, PC1 y PC2 explican aproximadamente el 98.7% de la varianza total, lo que indica que la información principal de los datos puede resumirse eficazmente en dos dimensiones.

El análisis de las cargas (loadings) reveló que:

- PC1 está fuertemente influido por la edad al diagnóstico, lo que sugiere que esta variable es la principal responsable de la variabilidad entre los pacientes.
- PC2 está dominado por el número de ganglios linfáticos positivos, seguido por una menor contribución de variables relacionadas con la invasión tumoral (linfática, venosa y perineural).

```r
# Realizamos agrupamiento jerárquico aglomerativo
dist_matrix <- dist(df_numeric)
# Aplicamos agrupamiento jerárquico aglomerativo con el método de Ward
hc_aglomerativo <- hclust(dist_matrix, method = "ward.D2")
```

```r
# Reprsentamos el dendrograma
plot(hc_aglomerativo, main = "Dendrograma de Agrupamiento Jerárquico
Aglomerativo", xlab = "Pacientes")
```



Sección 5: Visualización con Shiny

La aplicación Shiny fue desarrollada con el objetivo de permitir un análisis interactivo de datos clínicos de pacientes con cáncer gastrointestinal, utilizando un archivo CSV cargado por el usuario y un enfoque exploratorio y visual basado en el archivo `.Rmd`. El archivo `gastro_app.R` incluye el código fuente de la aplicación (el cual se adjunta en la entrega de la PEC4) y a continuación se muestran algunas capturas de pantalla para visualizar su funcionamento.

En términos generales, la aplicación incorpora transformaciones clave sobre el dataset, incluyendo el cálculo del tiempo de supervivencia (os_time), el evento de muerte (os_event) y el índice de masa corporal (IMC), a partir de variables como peso y altura. Además, se generan nuevas variables categóricas para la edad (age_category) y el IMC (imc_category), lo que facilita el análisis comparativo entre distintos grupos de pacientes.

La aplicación contiene un panel lateral con selectores que permiten al usuario elegir variables para tres secciones principales:

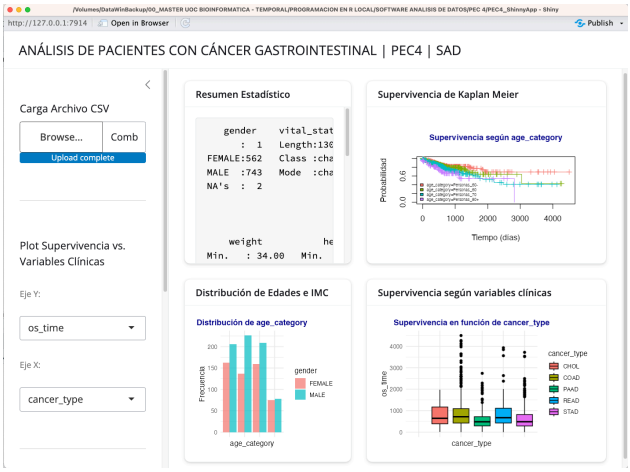


Figure 1: Vista general de la aplicación después de cargar el archivo de datos CSV

- **Supervivencia según variables clínicas**: Aquí se muestra un gráfico de tipo boxplot, en el cual se representa la relación entre una variable categórica (como tipo de cáncer o estadio tumoral) y una variable cuantitativa (por ejemplo, tiempo de supervivencia o edad).

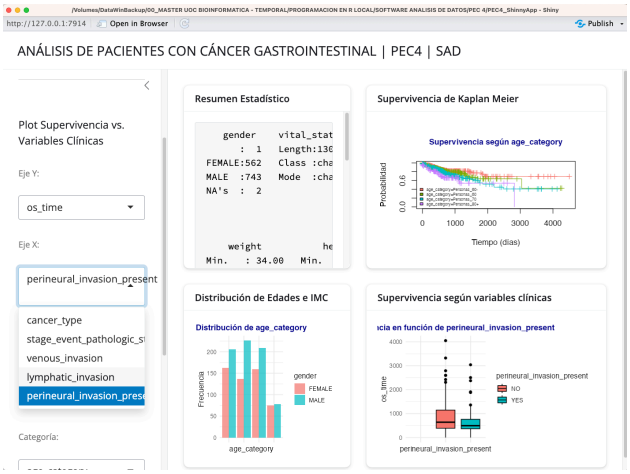


Figure 2: Supervivencia según variables clínicas

- **Distribución de Edades e IMC**: A través de un gráfico de barras, se analiza la distribución de frecuencias de variables como la edad y el IMC, estratificadas por género.

- **Análisis de Supervivencia Kaplan-Meier**: Utilizando la librería survival, se ajusta un modelo Kaplan-Meier para evaluar la probabilidad de supervivencia a lo largo del tiempo, según variables seleccionadas como género o grupo etario.

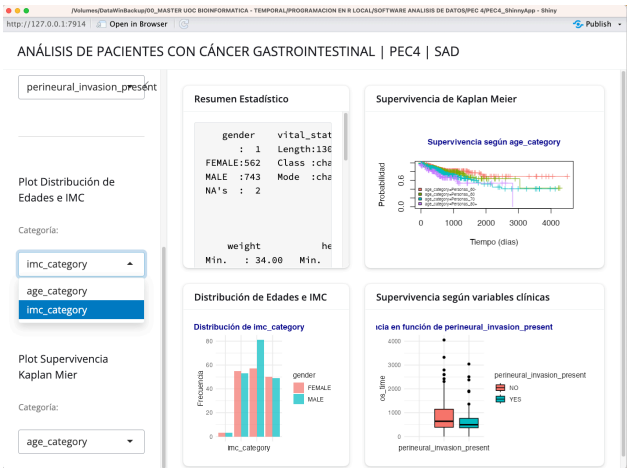


Figure 3: Distribución de Edades e IMC

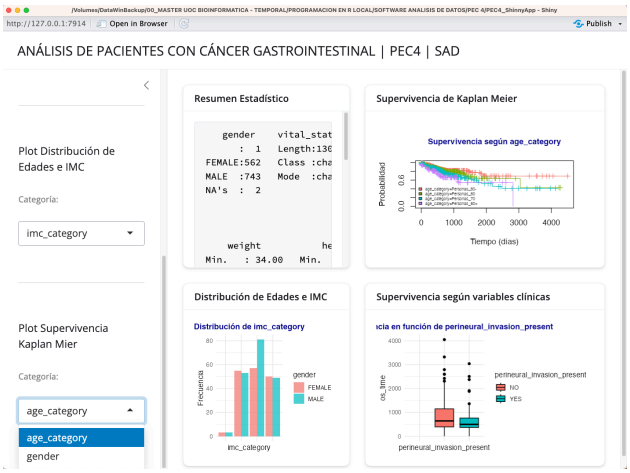


Figure 4: Análisis de Supervivencia Kaplan-Meier

Sección 6: Conclusiones

El objetivo del proyecto fue analizar datos clínicos de pacientes con cáncer gastrointestinal obtenidos del repositorio público TCGA (The Cancer Genome Atlas), para estudiar posibles factores clínicos que influyen en la supervivencia de los pacientes. Del dataset de más de 1.300 casos: 1) Seleccionamos variables relevantes, limpiamos y trasformamos datos, 2) Creamos gráficos exploratorios y realizamos análisis estadísticos, 3) Implementamos análisis de supervivencia de Kaplan-Mier, 4) Entrenamos tanto un modelo supervisado (SVM) para predecir la supervivencia como uno no supervisado (PCA y clustering jerárquico) para explorar agrupamientos naturales entre pacientes e 5) Implementamos una aplicación en Shiny para facilitar la visualización interactiva de estos datos.

A continuación se describen los principales hallazgos del análisis del TCGA dataset. Asimismo, se discuten limitaciones y áreas de oportunidad y se cierra con un comentario sobre la experiencia del trabajo en equipo para el desarrollo de este proyecto.

Sobre los hallazgos más interesantes del análisis

- Los resultados obtenidos con las pruebas paramétricas `ANOVA` y `TukeyHSD`, así como con las pruebas no paramétricas `Kruskal-Wallis` y `Dunn` confirmaron que existen diferencias estadísticamente significativas entre los tipos de cáncer gastrointestinal y los días de supervivencia para los siguientes grupos: Páncreas-Colon, Páncreas-Recto, Estómago-Colon y Stomágo-Recto.
- Los gráficos de Kaplan-Mier mostraron que las mujeres presentan una mayor supervivencia que los hombres ya que alrededor del 65% sobreviven más allá de 1.800 días, frente a la supervivencia del 35% de los hombres a los 3.000 días. El test de log-rank ($p = 0.04$) confirma que estas diferencias son estadísticamente significativas, sugiriendo una evolución clínica más favorable para las mujeres.
- Respecto a la edad, los pacientes más jóvenes presentan curvas de supervivencia más altas y sostenidas, mientras que las de los mayores de 70 y 80 años caen más rápido. El test de log-rank ($p = 0.0005$) presenta evidencia de que existen diferencias significativas entre grupos etarios. Esto es consistente con el hecho de que la edad avanzada se puede asociar a una mayor mortalidad y menor tolerancia a los tratamientos.
- Finalmente al analizar la probabilidad condicional, se observó que el 55.22% de los pacientes con invasión perineural presentaban un estadio patológico avanzado (tipo III o superior) al momento del diagnóstico. Esto sugiere una fuerte asociación entre la presencia de invasión perineural y un mayor grado de progresión tumoral.

Sobre las limitaciones de los datos y posibles mejoras o extensiones

Una de las principales limitaciones del análisis fue la alta proporción de datos faltantes, especialmente en variables categóricas como `venous_invasion`, `lymphatic_invasion` y `perineural_invasion_present`, así como en algunas variables continuas como `days_to_death`, `weight` y `height`. Esta situación redujo el tamaño efectivo de la muestra. Además, dado que se trata de un estudio observacional retrospectivo, es decir que los datos provienen de un repositorio existente y no de un diseño experimental controlado, fue posible identificar asociaciones, pero no establecer relaciones causales. Como mejora, podríamos aplicar técnicas de imputación para manejar los datos faltantes y validar los resultados con cohortes externas. Además, algunos estudios de ésta índole también incorporan datos genómicos para enriquecer el análisis y aportar una visión más integral de los factores que influyen en la supervivencia.

Sobre el trabajo en equipo y valoración del proyecto

Comentario de Salomón. El trabajo en equipo con Sefora fue clave para seleccionar el dataset, gracias a su experiencia en investigación oncológica. Esto nos ayudó a definir preguntas clave sobre cómo ciertas variables clínicas podrían afectar la supervivencia de los pacientes con cáncer gastrointestinal. Disfruté mucho de este proyecto porque me permitió aterrizar conceptos teóricos de una manera práctica con un

dataset real. Creo que la colaboración fue efectiva pues dividimos tareas, mantuvimos buena comunicación y revisamos mutuamente nuestro trabajo.

Comentario de Sefora. El proyecto nos permitió aplicar múltiples herramientas de análisis de datos en R y profundizar en el estudio de la supervivencia en cáncer gastrointestinal que es el tema de mi investigación en el Vall d'Hebron Instituto de Oncología. Espero poder aplicar las mismas herramientas para analizar los datos de los pacientes de mi centro y avanzar en el conocimiento de estas devastadoras enfermedades. Fue un placer trabajar con Salomon, ya que tenemos enfoques y perspectivas complementarios que nos permitieron mejorar mutuamente nuestro trabajo. Tuvimos una buena comunicación y nos repartimos las tareas a partes iguales.